U.S. DEPARTMENT OF COMMERCE
Patent and Trademark Office

SEARCH REQUEST FORM

4/26/02

Requestor's Name: Rita Mitra

Serial Number: 09-556,246

Date: 4-25-02 Phone: 605-1211 Art Unit: 1653

Rm 9803 MB 9801

Search Topic:

Please write a detailed statement of search topic. Describe specifically as possible the subject matter to be searched. Define any terms that may have a special meaning. Give examples or relevant citations, authors, keywords, etc., if known. For sequences, please attach a copy of the sequence. You may include a copy of the broadest and/or most relevant claim(s).

Please search

1. AA1 Seq 1, aa 1-24 plus aa 200-1404 (delete aa 25-199)
2. AA2 Seq 1, aa 1-25 plus 67-1404 (delete aa 26-66)
3. AA3 Seq 1, aa 1-25 plus aa 67-106 plus aa 200-1404 (delete aa 26-66 and 107-199)
4. aa 200-1140 of Seq 1
5. Seq 1 aa 25-end (delete aa 1-24)
6. AA4 Seq 1 aa 1-66 plus 105-end (delete aa 67-104)
7. AA5 Seq 1, aa 1-156 plus aa 200-1404 (delete aa 157-199)
8. AA6 Seq 1, aa 1-106 plus aa 200-1404 (delde aa 107-199)
9. Seq 1, aa 200-1167
10. Seq 1 aa 200-1212
11. Seq 1, aa 200-1263

Point of Contact:
Toby Port
Technical Info. Specialist
CM1 6A04
703-308-3534

STAFF USE ONLY

Date completed: 4/29

Searcher: T. Port + B. O'Bryan

Terminal time: 120

Elapsed time: 20

CPU time:

Total time:

Number of Searches:

Number of Databases:

Search Site

- STIC
- CM-1
- Pre-S

Type of Search

- N.A. Sequence
- 10 A.A. Sequence
- Structure
- Bibliographic

Vendors

- IG Suite
- STN
- Dialog
- APS
- Geninfo
- SDC
- DARC/Questel
- T6C+C6 Other

BEST AVAILABLE COPY

```
RA   Julien J.-P., Cote F., Beaudet L., Sidky M., Flavell D., Grosveld F.,
RA   Mushynski W.;
RT   "Sequence and structure of the mouse gene coding for the largest
RT   neurofilament subunit.";
RL   Gene 68:307-314(1988).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=89089138; PubMed=3145094;
RA   Shneidman P.S., Carden M.J., Lees J.F., Lazzarini R.A.;
RT   "The structure of the largest murine neurofilament protein (NF-H) as
RT   revealed by cDNA and genomic sequences.";
RL   Brain Res. 464:217-231(1988).
RN   [3]
RP   SEQUENCE FROM N.A.
RC   STRAIN=SWISS WEBSTER; TISSUE=Brain;
RA   Carden M.J.;
RL   Submitted (MAR-1994) to the EMBL/GenBank/DDBJ databases.
CC   -!- FUNCTION: NEUROFILAMENTS USUALLY CONTAIN THREE IF PROTEINS: L, M,
CC       AND H WHICH ARE INVOLVED IN THE MAINTENANCE OF NEURONAL CALIBER.
CC       NF-H HAS AN IMPORTANT FUNCTION IN MATURE AXONS THAT IS NOT
CC       SUBSERVED BY THE TWO SMALLER NF PROTEINS.
CC   -!- PTM: THERE ARE A NUMBER OF REPEATS OF THE TRIPEPTIDE K-S-P, NFH IS
CC       PHOSPHORYLATED ON A NUMBER OF THE SERINES IN THIS MOTIF. IT IS
CC       THOUGHT THAT PHOSPHORYLATION OF NFH RESULTS IN THE FORMATION OF
CC       INTERFILAMENT CROSS BRIDGES THAT ARE IMPORTANT IN THE MAINTENANCE
CC       OF AXONAL CALIBER.
CC   -!- PTM: PHOSPHORYLATION SEEMS TO PLAY A MAJOR ROLE IN THE FUNCTIONING
CC       OF THE LARGER NEUROFILAMENT POLYPEPTIDES (NF-M AND NF-H), THE
CC       LEVELS OF PHOSPHORYLATION BEING ALTERED DEVELOPMENTALLY AND
CC       COINCIDENT WITH A CHANGE IN THE NEUROFILAMENT FUNCTION.
CC   -!- SIMILARITY: BELONGS TO THE INTERMEDIATE FILAMENT FAMILY.
CC   -!- CAUTION: REF.2 SEQUENCE DIFFERS FROM THAT SHOWN IN POSITIONS 534
CC       TO 716 AND IS SHORTER DUE TO FRAMESHIFTS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M24496; AAA39813.1; -.
DR   EMBL; M23349; AAA39813.1; JOINED.
DR   EMBL; M24494; AAA39813.1; JOINED.
DR   EMBL; M24495; AAA39813.1; JOINED.
DR   EMBL; M35131; AAA39809.1; ALT_FRAME.
DR   EMBL; Z31012; CAA83229.1; -.
DR   PIR; JT0368; QFMSH.
DR   PIR; A43778; A43778.
DR   MGD; MGI:97309; Nfh.
DR   InterPro; IPR001664; IF.
DR   Pfam; PF00038; filament; 1.
DR   PROSITE; PS00226; IF; 1.
KW   Intermediate filament; Coiled coil; Neurone; Phosphorylation;
KW   Repeat.
FT   DOMAIN        1     97       HEAD.
FT   DOMAIN       98    408       ROD.
FT   DOMAIN      409   1087       TAIL.
FT   DOMAIN      436    517       GLU-RICH (ACIDIC).
FT   DOMAIN      519    886       50 X 6 AA TANDEM REPEATS OF K-S-P-A-E-A.
FT   DOMAIN      887   1087       GLU/LYS-RICH.
FT   DOMAIN       98    129       COIL 1A.
FT   DOMAIN      130    141       LINKER 1.
FT   DOMAIN      142    239       COIL 1B.
FT   DOMAIN      240    261       LINKER 12.
FT   DOMAIN      262    283       COIL 2A.
FT   DOMAIN      284    287       LINKER 2.
FT   DOMAIN      288    408       COIL 2B.
FT   CONFLICT    133    133       K -> QA (IN REF. 2 AND 3).
FT   CONFLICT    199    199       A -> AR (IN REF. 2 AND 3).
FT   CONFLICT    281    281       S -> T (IN REF. 2 AND 3).
FT   CONFLICT    492    492       L -> G (IN REF. 2 AND 3).
FT   CONFLICT    551    551       P -> PREAKSP (IN REF. 3).
FT   CONFLICT    689    712       MISSING (IN REF. 3).
FT   CONFLICT    714    714       G -> A (IN REF. 3).
FT   CONFLICT    814    814       V -> M (IN REF. 2 AND 3).
FT   CONFLICT    843    843       T -> N (IN REF. 2 AND 3).
SQ   SEQUENCE   1087 AA;  116612 MW;  57BAC76A38ED1CB9 CRC64;
```

```
  Query Match             7.6%;  Score 551;  DB 1;  Length 1087;
  Best Local Similarity   26.8%;  Pred. No. 5.8e-15;
  Matches  234;  Conservative   96;  Mismatches  358;  Indels  186;  Gaps   43;

Qy     105 KKTKKVIESE--EITEEHSVS-ENQESSSSSSSSSSSSTIWKIKSSKNSAANRELQKKLK 161
           | ||: :| :  |: : |     : : :     |   :| |::     :|  || | |
Db     331 KSTKESLERQRSELEDRHQADIASYQDAIQQLDSELRNTKWEM-----AAQLREYQDLLN 385

Qy     162 VK-------------DNKKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNK 207
           ||             : :: |    |:|         |  |  |: |  ||  | |
Db     386 VKMALDIEIAAYRKLLEGEECRIGFGPSP---------FSLTEGLPKI--PSIST--HIK 432

Qy     208 VSTSPKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTT 267
           | :   |  :           : ||| :   :   :   '|  |  :    |:|
Db     433 VKSEEMIKVVE-----------KSEKETVIVEGQTEEIRVTEGVTEEEDKEAQGQEGEE 480

Qy     268 SAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKE 327
           : :  :  |: |   :| ::  | |  : ::  |  | :  :  :|   | ||    :
Db     481 AEEGEEKEEEELAAATSPPAEEAASPEKETKSRVKEEAKSPGEAKSPGEAKSPA----EA 536

Qy     328 PTPTTIKSAPTTPKEPAPTTTKSAPTTPKEP-APTTTKEPAPTTPKEPAPTTTKEP---- 382
           :|   || |   | |    : : | :| || :|   | ||   || || | | |
Db     537 KSPGEAKS-PGEAKSPGEAKSPAEPKSPAEPKSPAEAKSPA--EPKSPA--TVKSPGEAK 591

Qy     383 APTTTKSAPTTPKEPA----PTTPKKPA----PTTPKEPA----PTTPKEP----TPTTP 426
           :|:  || |   | ||    |   | ||    |  | ||    | | | |    :|:
Db     592 SPSEAKS-PAEAKSPAEAKSPAEAKSPAEAKSPAEAKSPAEAKSPATVKSPGEAKSPSEA 650

Qy     427 KEPAPTTKEPAPTTPKEPA----PTAPKKP----APTTPKEPAPTTPKEPAPTTTKEPSP 478
           | ||   :  :|   | ||    |   | |     :|  || ||    | ||    : ||
Db     651 KSPA---EAKSPAEAKSPAEAKSPAEVKSPGEAKSPAEPKSPAEA--KSPA----EVKSP 701

Qy     479 TTPKEPA----PTTTKSAPTTTKEPAPTTTKSAPTTPKE-PSPTTTKEPA----PTTPKE 529
           | ||    |   || |   | ||   : :|  :| |  ||   | ||    |   |
Db     702 AEAKSPAEVKSPGEAKS-PAAVKSPAEAKSPAAVKSPGEAKSPGEAKSPAEAKSPAEAKS 760

Qy     530 PAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPT 589
           |      : | |  || | :  :  :|   | |      | ||      | |  :| |
Db     761 PIEVKSPEKAKTPVKEGAKSPAEAKSPEKAKSPVKEDIKPPAEAKSPEKA-KSPVKEGAK 819

Qy     590 TPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEE-PAPTTPKAAAPNTPKEPAPTT-- 646
           |||  |   : |   || :  | | :  |  ||:  :|   || :|   || | |:
Db     820 PPEKAKPLDVKSPEAQTPVQEEATVPTDIRP--PEQVKSPAKEKAKSPE--KEEAKTSEK 875

Qy     647 --PKEPAPTTP-------KEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAP 697
             ||:    :|       |||     :|     ||| |  : |: ||    ||  :|
Db     876 VAPKKEEVKSPVKEEVKAKEPPKKVEEEKTLPTPKTEAKESKKDEAPKEAPKPKVEEKKE 935

Qy     698 TTTKEPTSTTSD--KPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTT 755
           | |::|  :|::  |       |    :  : ||    || |   ||    ||  |   |
Db     936 TPTEKPKDSTAEAKKEEAGEKKKAVASEEETPAKLGVKEEA--KPKEKTETTKTEAEDTK 993

Qy     756 PKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETPPPT 815
           |:|:     || |: |                          ||  |  |:|
Db     994 AKEPS----KPTETEKP-------------------KKEEMPAAPEK------------ 1017

Qy     816 TSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTT 875
             | :| ||  : |:|           || :|   ||      :| |  :||:
Db    1018 ------KDTKEEKTTESRKPEEK---------PK-MEAKVKEDDKSLSKEP--SKPKTEK 1059

Qy     876 TAKDKTTERDLRTTPETTTAAPKMTKETATTTEK 909
             |  :|::     ||       | |::  ||  ||
Db    1060 AEKSSSTDQKESQPPE------KTTEDKATKGEK 1087
```

```
Qy     513 EPSPTTTKEPAPTTPKEPAPTTPKKPA-----PTTPKEPAPTTPKEPAPTTTKKPAPTAP 567
           |    :|:|  |||| |    : :||:     || | |   |: :|   :| | |
Db     702 EKPTISTEE--PTTPTEETTISTEKPSIPMEKPTLPTEETTTSVEETTISTEKLTIPM-- 757

Qy     568 KEPAPTTPKETAPT-----TPKKLTPTTPEKLA-----PTTPEKPAPTTPEELAPTTPEE 617
           ::|  :| | | ||     :|:||| | |||      || | :    : |:|  | | |
Db     758 EKPTISTEKPTIPTEKPTISPEKLTIPT-EKLTIPTEKPTIPIEETTISTEKL--TIPTE 814

Qy     618 PTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPT 677
             :||:|  :| |           || | |  || | |       ||  :| | | ||
Db     815 KPTISPEKPTISTEK----------PTIPTE-KPTIPTE-------ETTISTEKLTIPT 855

Qy     678 TLKEPAPTTPKKPAPKELAPTTTKEPTSTT------SDKPAPTTPKGTAPTTPKEPAPTT 731
             ::|  :  |   | |   :|::||  |      ::||   | | | ||  ::
Db     856 --EKPTISPEKLTIPTEKPTISTEKPTIPTEKLTIPTEKPTIPTEKPTIPT--EKLTALR 911

Qy     732 PKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKET--A 789
           |  |:||   | |   :   ||:||      :  |||   :||    | |    |:
Db     912 PPHPSPTA-TGLAALVMSPHAPSTPMTSV---ILGTTTTSRSSTGMSCP-PNARYESCAC 966

Qy     790 PTTPKEPAPT 799
           | : | | |:
Db     967 PASCKSPRPS 976


RESULT   7
EXTN_TOBAC
ID   EXTN_TOBAC     STANDARD;      PRT;   620 AA.
AC   P13983;
DT   01-JAN-1990 (Rel. 13, Created)
DT   01-JAN-1990 (Rel. 13, Last sequence update)
DT   01-MAR-1992 (Rel. 21, Last annotation update)
DE   EXTENSIN PRECURSOR (CELL WALL HYDROXYPROLINE-RICH GLYCOPROTEIN).
GN   HRGPNT3.
OS   Nicotiana tabacum (Common tobacco).
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC   Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots;
OC   Asteridae; euasterids I; Solanales; Solanaceae; Nicotiana.
OX   NCBI_TaxID=4097;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=CV. XANTHI; TISSUE=Leaf;
RX   MEDLINE=90128263; PubMed=2612909;
RA   Keller B., Lamb C.J.;
RT   "Specific expression of a novel cell wall hydroxyproline-rich
RT   glycoprotein gene in lateral root initiation.";
RL   Genes Dev. 3:1639-1646(1989).
CC   -!- FUNCTION: HAS A SPECIALIZED STRUCTURAL FUNCTION, POSSIBLY IN
CC       THE MECHANICAL PENETRATION OF THE CORTEX AND EPIDERMIS OF THE
CC       MAIN ROOT.
CC   -!- SUBCELLULAR LOCATION: EXTRACELLULAR MATRIX.
CC   -!- PTM: EXTENSINS CONTAIN A CHARACTERISTIC REPEAT OF THE PENTAPEPTIDE
CC       SER-PRO(4). THE PROLINE RESIDUE IS HYDROXYLATED AND THEN
CC       GLYCOSYLATED.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X13885; CAA32090.1; -.
DR   PIR; S06733; S06733.
KW   Repeat; Cell wall; Glycoprotein; Signal; Structural protein;
KW   Hydroxylation.
FT   SIGNAL        1      ?
FT   CHAIN         ?    620       EXTENSIN.
FT   REPEAT       70     73       H-A-P-P.
FT   REPEAT      148    151       H-A-P-P.
FT   DOMAIN      229    242       2 X 7 AA TANDEM REPEATS OF T-H-R-H-A-P-P.
FT   REPEAT      229    235       1.
FT   REPEAT      236    242       2.
FT   DOMAIN      205    620       CONTAINS THE SER-PRO(4) REPEATS.
FT   DOMAIN      499    600       3 X APPROXIMATE TANDEM REPEATS.
SQ   SEQUENCE   620 AA;  65406 MW;  641DD2278AB28524 CRC64;


  Query Match             7.6%;  Score 551;  DB 1;  Length 620;
  Best Local Similarity   27.2%;  Pred. No. 3.6e-15;
  Matches  170;  Conservative   64;  Mismatches  304;  Indels   88;  Gaps   24;

Qy     297 AETTTK-GPALTTP--KEPTPTT-----PKEPASTTPKE---PTPTTIKSAPTTPKEPAP 345
           || ||: | | |   :| |::      | | :| |     |:|   : ||     | |
Db      24 AEATTQYGGYLPPPVTSQPPPSSIGLSPPSAPTTTPPSRGHVPSP---RHAPPRHAYPPP 80

Qy     346 TTTKSAPTTPKEP------APTTTKEPAPT---TPKEPAPTTTKEPAPTTTKSAPTTPKE 396
           :    |:    |       |:    | |:   :|  | |:    |          |:  :
Db      81 SHGHLPPSVGGPPPHRGHLPPSRGFNPPPSPVISPSHPPPSYGAPPPSHGPGHLPSHGQR 140

Qy     397 PAPTTPKKPAPT---TPKEPAPTTPKEPTPTTPKEPAPTTK--EPAPTTPKEPAPTAPKK 451
           |       |:|:    |     | |:   | : : |:| ::   | | |  :| ||
Db     141 P-------PSPSHGHAPPSGGHTPPRGQHPPSHRRPSPPSRHGHPPPPTYAQPPPTPIYS 193

Qy     452 PAPTTPKEPAPTTPKEPAPTTTKEPSPTTPK---EPAPTTTKSAPTTTKEPAPTTTKS-- 506
           |:|    :| ||    | | |  :|:|: |    :| | | : || | :   ||   |
Db     194 PSPQV--QPPPT--YSPPPPTHVQPTPSPPSRGHQPQPPTHRHAPPTHRHAPPTHQPSPL 249

Qy     507 ---APTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPT---TPKEPAPTTTK 560
             |:  ::| | |   | |   : | |:    | | |   | |:   :|  || : :
Db     250 RHLPPSPRRQPQPPTYSPPPPAYAQSPQPSPTYSPPPPTYSPPPPSPIYSPPPPAYSPSP 309

Qy     561 KPAPT---APKEPAPTTPKETAPTTPKKL-TPTT------PEKLAPTTPEKPAPTTPEEL 610
           | ||   :|  || : |   :|  |  |  |::       |   :|  |   :|  | |
Db     310 PPTPTPTFSPPPPAYSPPPTYSPPPPTYLPLPSSPIYSPPPPVYSPPPPPSYSPPPPTYL 369

Qy     611 APTTPEEPTPTTPEEPAPTTPKAAAP---NTPKEPAPTTPKEPAPTTPKEPAPTTPKETA 667
           |  |  | | :   | ||  ::  |    :|  |||  |   | ||     | | |  :
Db     370 PPPPPSSPPPPSFSPPPPTYEQSPPPPPAYSPPLPAPPTYSPPPPT--YSPPPPTYAQPP 427

Qy     668 PTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEP 727
           |       | |   | |     | |    |  |  |       :| |  |  : |     |
Db     428 P------LPPTYSPPPPAYSPPPPPTYSPPPPTYSPPPPAYAQPPPPPPTYSPPPPAYSP 481

Qy     728 APTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKE 787
           | :|   |          | |    | |     | |    |   : || |    |:| |
Db     482 PPPSPIYSPPPPQVQPLPPTFSPPPPRRIHLPPPPHRQP---RPPTPTYGQPPSPPTFSP 538

Qy     788 TAPTTPKEPAPT--TPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAE 845
           |     | |    |: | ||  : | | |      |    |   |: |  ||
Db     539 PPPRQIHSPPPPHWQPRTPTPTYGQPPSPPTFSAPPPRQIHSPPPPHRQPRPPTPTYGQP 598

Qy     846 PTPKALENSPKEPGVPTTKTPAATKP 871
           |:|           ||| :| :  |
Db     599 PSP-----------PTTYSPPSPPP 612


RESULT   8
NFH_MOUSE
ID   NFH_MOUSE      STANDARD;      PRT;  1087 AA.
AC   P19246; Q61959;
DT   01-NOV-1990 (Rel. 16, Created)
DT   01-AUG-1992 (Rel. 23, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   NEUROFILAMENT TRIPLET H PROTEIN (200 KDA NEUROFILAMENT PROTEIN)
DE   (NEUROFILAMENT HEAVY POLYPEPTIDE) (NF-H).
GN   NEFH OR NFH.
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=89121513; PubMed=3220257;
```

```
Qy     850 ALENSPK----EPGVPTT--KTPA-----ATKPEMTTTAKDKTTERDLRTTPETTTAAPK 898
           |   :|     :|  | |    |:       ||    | | | | :   | | |  | |
Db     689 AKPTNPSTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTYPSTYKAKP- 747

Qy     899 MTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTE 958
           | :   |   | ::|        :   |    |    | |: :: |     |||   |
Db     748 -TYKAKPTYPPTYKAK-------PSYPPTYKPKPSYPPTYKSKSIYPSSYKPKKTYPPT- 798

Qy     959 IMNKPEETAKPKDRATNSKATTPKPQKPTKAPKK---PTSTKKPKTMPRVRKPKTTPTP- 1014
             ||: |  |  :        |||  |     |   |:: |   : |   | ||:  |
Db     799 --YKPKLTYPPTYK--------PKPSYPPSYKPKITYPSTYKLKPSYPPTYKSKTSYPPT 848

Qy    1015 -----------RKMTSTMPELNPTSR 1029
                      :  ||  |   ||:|
Db     849 YNKKISYPSSYKAKTSYPPAYKPTNR 874
```

```
RESULT   6
ZAN_HUMAN
ID   ZAN_HUMAN      STANDARD;      PRT;  2700 AA.
AC   Q9Y493; O00218;
DT   20-AUG-2001 (Rel. 40, Created)
DT   20-AUG-2001 (Rel. 40, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ZONADHESIN (FRAGMENT).
GN   ZAN.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE OF 1-2379 FROM N.A.
RX   MEDLINE=99018118; PubMed=9799793;
RA   Glockner G., Scherer S., Schattevoy R., Boright A., Weber J.,
RA   Tsui L.C., Rosenthal A.;
RT   "Large-scale sequencing of two regions in human chromosome 7q22:
RT   analysis of 650 kb of genomic sequence around the EPO and CUTL1 loci
RT   reveals 17 genes.";
RL   Genome Res. 8:1060-1073(1998).
RN   [2]
RP   SEQUENCE OF 2338-2700 FROM N.A.
RC   TISSUE=Testis;
RX   MEDLINE=97271566; PubMed=9126492;
RA   Gao Z., Harumi T., Garbers D.L.;
RT   "Chromosome localization of the mouse zonadhesin gene and the human
RT   zonadhesin gene (ZAN).";
RL   Genomics 41:119-122(1997).
CC   -!- FUNCTION: BINDS IN A SPECIES-SPECIFIC MANNER TO THE ZONA PELLUCIDA
CC       OF THE EGG. MAY BE INVOLVED IN GAMETE RECOGNITION AND/OR
CC       SIGNALING.
CC   -!- SUBUNIT: PROBABLY FORMS COVALENT OLIGOMERS.
CC   -!- SUBCELLULAR LOCATION: TYPE I MEMBRANE PROTEIN, EXCLUSIVELY ON THE
CC       APICAL REGION OF THE SPERM HEAD (BY SIMILARITY).
CC   -!- TISSUE SPECIFICITY: IN TESTIS; PRIMARILY IN HAPLOID SPERMATIDS.
CC   -!- DOMAIN: THE MAM DOMAINS PROBABLY MEDIATE SPERM ADHESION TO THE
CC       ZONA PELLUCIDA.
CC   -!- DOMAIN: DURING SPERM MIGRATION THROUGH THE REPRODUCTIVE TRACTS,
CC       THE MUCIN-LIKE DOMAIN MIGHT INHIBIT INAPPROPRIATE TRAPPING OF
CC       SPERMATOZOA OR PROMOTING ADHESION TO THE OVIDUCTAL ISTHMUS.
CC   -!- DOMAIN: THE VWFD DOMAINS 2 AND 3 MAY MEDIATE COVALENT
CC       OLIGOMERIZATION (BY SIMILARITY TO HUMAN INTESTINAL MUCIN MUC2).
CC   -!- SIMILARITY: CONTAINS 3.5 MAM DOMAINS.
CC   -!- SIMILARITY: CONTAINS 4.5 VWFD DOMAINS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; AF053356; AAC78790.1; -.
DR   EMBL; U83191; AAC51208.1; -.
DR   MIM; 602372; -.
DR   InterPro; IPR000561; EGF-like.
DR   InterPro; IPR000998; MAM.
DR   InterPro; IPR002965; P_rich_extensn.
DR   InterPro; IPR002919; TIL.
DR   InterPro; IPR003328; TILa.
DR   InterPro; IPR001007; VWFC.
DR   InterPro; IPR001846; Vwd.
DR   Pfam; PF00629; MAM; 3.
DR   Pfam; PF01826; TIL; 4.
DR   Pfam; PF02345; TILa; 4.
DR   Pfam; PF00094; vwd; 4.
DR   PRINTS; PR01217; PRICHEXTENSN.
DR   SMART; SM00137; MAM; 2.
DR   SMART; SM00214; VWC; 1.
DR   SMART; SM00011; VWC_def; 3.
DR   SMART; SM00216; VWD; 3.
DR   PROSITE; PS01186; EGF_2; 3.
DR   PROSITE; PS00740; MAM_1; 1.
DR   PROSITE; PS50060; MAM_2; 4.
KW   Glycoprotein; Transmembrane; Cell adhesion; Repeat.
FT   NON_TER       1      1
FT   DOMAIN       <1    109       MAM 1.
FT   DOMAIN      112    136       MAM 2 (PARTIAL).
FT   DOMAIN      161    326       MAM 3.
FT   DOMAIN      322    446       MAM 4.
FT   DOMAIN      483    951       66 X HEPTAPEPTIDE REPEATS (APPROXIMATE)
FT                                (MUCIN-LIKE DOMAIN).
FT   DOMAIN      953   1065       VWFD 1 (PARTIAL).
FT   DOMAIN     1066   1454       VWFD 2.
FT   DOMAIN     1455   1861       VWFD 3.
FT   DOMAIN     1862   2292       VWFD 4.
FT   DOMAIN     2293   2684       VWFD 5.
FT   DOMAIN        ?      ?       EGF-LIKE.
FT   CARBOHYD     74     74       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    403    403       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1023   1023       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1099   1099       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1618   1618       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1737   1737       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1832   1832       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1878   1878       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   2136   2136       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   2505   2505       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CONFLICT   2374   2379       NNQKMA -> RAGPGP (IN REF. 1).
FT   NON_TER    2700   2700
SQ   SEQUENCE   2700 AA;  293013 MW;  80E60CC0B12277B1 CRC64;
```

```
  Query Match             7.6%;  Score 555.5;  DB 1;  Length 2700;
  Best Local Similarity   32.6%;  Pred. No. 8.3e-15;
  Matches  199;  Conservative   64;  Mismatches  240;  Indels  107;  Gaps   32;

Qy     219 PINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKT 278
           |:   | ||| |  |      |   |||  |:  | : |             | | ||||
Db     445 PVKVLPELPPVSPVSS----TGPSETTGLTENPTISTK------------KPTVSIEKP 487

Qy     279 SAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPT-----PTTI 333
           |     ||      | |    |: | ::| |   || | |  :   ::||     || :
Db     488 SVTTEKPT-----VPKEKPTIPTEKPTISTEK---PTIPSEKPNMPSEKPTIPSEKPTIL 539

Qy     334 KSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPA-PTTTKSAPT 392
             || | |  ||      ||   |     |:|  |||| |    |  :||  ||   | ||
Db     540 TEKPTIPSE-KPTIPSEKPTISTEKPTVPTEE--PTTPTEETTTYMEEPVIPTEKPSIPT 596

Qy     393 TPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKP 452
             ::|:  | |   ||   |     : ::||     |   || |   || | | :  :|:|
Db     597 --EKPSIPTEK---PTISMEETIISTEKPTICPEKPTIPTEK---PTIPTEKSTISPEK- 647

Qy     453 APTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPK 512
           |||| |  || | |    :|::|:  | | |  :  |    | |   ||     || |
Db     648 -PTTPTE-KPTIPTEKPTISTEKPTIPTEK-PTISPEKLTIPTEKLTIPT---EKPTIPT 701
```

```
Qy    821 TPTTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEM 873
Db    701 APDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAP-----DTRPAPGSTAP-- 753

Qy    874 TTTAKDKTTERDLRTTPETT-------TAAPKMTKETATTTEKTTESKITATTTQVTSTT 926
Db    754 --PAHGVTSAPDTRPAPGSTAPPAHGVTSAP----DTRPAPGSTAPPAHGVTSAPDTRPA 807

Qy    927 TQDTTP--FKITTLKTTTLAPKVTTTKKTITTT--EIMNKPEETAKPKDRATNSKATTPK 982
Db    808 PGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPA 867

Qy    983 PQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPN 1042
Db    868 P--GSTAPPAHGVTSAPDTRP---APGSTAPPAHGVTSAPDTRPA------------PG 909

Qy   1043 QT-PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPML 1101
Db    910 STAPPAHGVTSAPDTRPAPGSTAPPAH----------GVTSAPDNRPALGSTA----PPV 955

Qy   1102 SDETNICNGKPVDGLTTLRNGTLVAFRGHYFWMLSPFSPPS 1142
Db    956 HNVTSASGSASGSASTLVHNGTSARATTTPASKSTPFSIPS 996
```

```
RESULT   5
FP1_MYTED
ID   FP1_MYTED      STANDARD;      PRT;   875 AA.
AC   Q25460;
DT   30-MAY-2000 (Rel. 39, Created)
DT   30-MAY-2000 (Rel. 39, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ADHESIVE PLAQUE MATRIX PROTEIN (POLYPHENOLIC ADHESIVE PROTEIN) (FOOT
DE   PROTEIN 1) (MEFP1) (FRAGMENT).
GN   FP1.
OS   Mytilus edulis (Blue mussel).
OC   Eukaryota; Metazoa; Mollusca; Bivalvia; Pteriomorphia; Mytiloida;
OC   Mytiloidea; Mytilidae; Mytilus.
OX   NCBI_TaxID=6550;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=91025829; PubMed=1367451;
RA   Filpula D.R., Lee S.M., Link R.P., Strausberg S.L., Strausberg R.L.;
RT   "Structural and functional repetition in a marine mussel adhesive
RT   protein.";
RL   Biotechnol. Prog. 6:171-177(1990).
RN   [2]
RP   PARTIAL SEQUENCE, AND POST-TRANSLATIONAL MODIFICATIONS.
RX   MEDLINE=83135732; PubMed=6298211;
RA   Waite J.H.;
RT   "Evidence for a repeating 3,4-dihydroxyphenylalanine- and
RT   hydroxyproline-containing decapeptide in the adhesive protein of the
RT   mussel, Mytilus edulis L.";
RL   J. Biol. Chem. 258:2911-2915(1983).
CC   -!- FUNCTION: PROVIDES ADHESIVENESS TO THE MUSSEL'S FOOT. MUSSELS
CC       PRODUCE ONE OF THE STRONGEST WATER INSOLUBLE GLUE. THE MUSSEL'S
CC       ADHESIVE IS A BUNDLE OF THREADS, CALLED A BYSSUS, FORMED BY A
CC       FIBROUS COLLAGENOUS CORE COATED WITH ADHESIVE PROTEINS.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- TISSUE SPECIFICITY: PRODUCED BY THE BYSSAL GLAND.
CC   -!- DOMAIN: ALMOST EXCLUSIVELY COMPOSED OF REPEATS OF A DECAPEPTIDE.
CC   -!- PTM: THE DECAPEPTIDE A-K-P-S-Y-P-P-T-Y-K IS POST-TRANSLATIONALLY
CC       MODIFIED AS FOLLOWING: THE SIXTH AND SEVENTH RESIDUES ARE
CC       HYDROXYLATED AND THE PENULTIMATE IS A 3,4-DIHYDROXYPHENYLALANINE
CC       (DOPA) DERIVED FROM TYROSINE.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
```

```
CC   ---------------------------------------------------------------------
DR   EMBL; X54422; CAA38294.1; -.
DR   InterPro; IPR002964; Adhesive_plaq.
DR   InterPro; IPR002965; P_rich_extensn.
DR   PRINTS; PR01216; ADHESIVEI.
DR   PRINTS; PR01217; PRICHEXTENSN.
KW   Repeat; Hydroxylation.
FT   NON_TER       1      1
FT   DOMAIN       67    870       TANDEM REPEATS OF Y-K-[PA]-K-[LP]-[ST]-Y-
FT                                 P-P-[ST].
SQ   SEQUENCE   875 AA;  100412 MW;  6EA85312748CAACE CRC64;
```

Query Match 7.7%; Score 559; DB 1; Length 875;
Best Local Similarity 26.9%; Pred. No. 2.4e-15;
Matches 281; Conservative 102; Mismatches 403; Indels 260; Gaps 56;

```
Qy     70 HNPTSPPSSKKAPPPSGASQTIKSTTKRSPKPPNKKKTKKVIESEEITEEHSVSENQESS 129
Db      3 HEPVYKP-------------VKTSYSAPYKPPTYQPLKKKVD-------------YRPT 35

Qy    130 SSSSSSSSSSTIWKIKSSKNSAANRELQKKLKVKDNKKNRTKKKPTPKPPVVDEAGSGLD 189
Db     36 KSYPPTYGSKTNYLPLAKKLSS-----YKPIKTTYNAK-------TNYPPVY-------- 75

Qy    190 NGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETK 249
Db     76 --------------------KPKMTYPPTYKPKPSYPP----------TYKSKPTYKPK 104

Qy    250 ETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTP 309
Db    105 IT-----------YPPTYKAKPSYPSSYKPKKTYPPTYK------PKLTY----PPTYKP 143

Qy    310 KEPTPTT----PKEPASTTPKEPTPTTIKSAPTTP-----KEPAPTTTKSAPTTPKEPAP 360
Db    144 KPSYPPTYKPKPSYPPSYKTKKTYPSSYKAKPSYPPTYKAKPSYPPTYKAKPSYP----P 199

Qy    361 TTTKEPA-PTTPKEPAPTTTKEPAPTTTKSAPTTPKEPA--PTTPKKPA--PTTPKEPA- 414
Db    200 TYKAKPTYKAKPTYPSTYKAKPSYPPTYKAKPTYKAKPSYPPTYKAKPSYPPTYKAKPSY 259

Qy    415 -PTTPKEPT-PTTPKEPAPTTKEPAPTTPKEPAPTAPKKPA--PTTPKEPA--PTTPKEP 468
Db    260 PPTYKAKPTYKAKPTYKAKPTYKAKPSYP----PTYKAKPSYPPTYKAKPSYPPTYKAKP 315

Qy    469 A--PTTTKEPS--PTTPKEPA-PTTTKSAPTTTKEPA-PTTTKSAPTTPKEPSPTTTKEP 522
Db    316 SYPPTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTYPSTYKAKPSYP----PTYKAKP 371

Qy    523 APTTPKEPAPTTPKKPA--PTTPKEPA--PTTPKEPAPTTTKKPAPTAPKEPAPTTPKET 578
Db    372 SYPPTYKAKPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKP-------T 424

Qy    579 APTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNT 638
Db    425 YPST-YKAKPSYP----PSYKAKPSYPPTYKAKPTYKAKPT-------YPSTYKA----K 468

Qy    639 PKEPAPTTPKEPAPTTPKE----PAPTTPKETAPTT--PKGTAPTTLKEPAPTTPKKPAP 692
Db    469 PSYPASYKAKPSYPPTYKSKSSYPSSYKPKKTYPPTYKPKLTYKPTYK-PKPSYPPSYKP 527

Qy    693 KELAPTTTKEPTSTTSDKPAPTTPKGTAPTT----PKEPAPTTPKEPAPTTPKGT---AP 745
Db    528 KTTYPPTYK--------------PKISYPPTYKAKPSYPATYKAKPSYPPTYKAKPSYPP 573

Qy    746 TTLKEPA--PTTPKKPA--PKELAPTTTKG-PTSTTSDKPAPTTPKETAPTTPKEPA--P 798
Db    574 TYKAKPSYPPTYKAKPSYKAKPTYPSTYKAKPSYPPTYKAKPSYP----PTYKAKPSYPP 629

Qy    799 TTPKKPA-PTT----PETPPPTTSEVSTPTTTKE----PTTIHKSPDESTPELSAEPTPK 849
Db    630 TYKAKPTYPSTYKAKPSYPPTYKPKISYPPTYKAKPSYPPT-YKAKPSYPPTYKAKPTYK 688
```

```
RT   "Preoperative diagnosis of thyroid papillary carcinoma by reverse
RT   transcriptase polymerase chain reaction of the MUC1 gene.";
RL   Int. J. Cancer 66:55-59(1996).
RN   [11]
RP   SEQUENCE OF 1-89 FROM N.A.
RC   TISSUE=Lung;
RX   MEDLINE=96181716; PubMed=8604237;
RA   Yu C.J., Yang P.C., Shew J.Y., Hong T.M., Yang S.C., Lee Y.C.,
RA   Lee L.N., Luh K.T., Wu C.W.;
RT   "Mucin mRNA expression in lung adenocarcinoma cell lines and
RT   tissues.";
RL   Oncology 53:118-126(1996).
RN   [12]
RP   SEQUENCE OF 1-46 FROM N.A.
RC   TISSUE=Breast carcinoma;
RA   Buluwela L., Liu Q., Luqmani Y.A., Gomm J.J., Coombes R.C.;
RL   Submitted (OCT-1992) to the EMBL/GenBank/DDBJ databases.
CC   -!- FUNCTION: DIRECT OR INDIRECT INTERACTION WITH ACTIN
CC       CYTOSKELETON.
CC   -!- SUBCELLULAR LOCATION: TYPE I MEMBRANE PROTEIN. A SECRETED FORM
CC       IS ALSO PRODUCED.
CC   -!- ALTERNATIVE PRODUCTS: VARIOUS VARIANTS ARE PRODUCED BY ALTERNATIVE
CC       SPLICING.
CC   -!- TISSUE SPECIFICITY: ABERRANTLY EXPRESSED IN HUMAN EPITHELIAL
CC       TUMORS, SUCH AS BREAST CANCER.
CC   -!- PTM: HIGHLY GLYCOSYLATED (N-AND O-LINKED CARBOHYDRATES AND SIALIC
CC       ACID).
CC   -!- POLYMORPHISM: THE NUMBER OF REPEATS IS HIGHLY POLYMORPHIC. IT
CC       VARIES FROM 21 TO 125 IN THE NORTHERN EUROPEAN POPULATION. THE
CC       MOST FREQUENT ALLELES CONTAINS 41 AND 85 REPEATS.
CC   -!- SIMILARITY: CONTAINS 1 SEA DOMAIN.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; J05582; AAA60019.1; -.
DR   EMBL; M32738; AAA35804.1; -.
DR   EMBL; M32739; AAA35806.1; -.
DR   EMBL; J05581; AAA59876.1; -.
DR   EMBL; M61170; AAB53150.1; -.
DR   EMBL; X52229; CAA36478.1; ALT_SEQ.
DR   EMBL; X52228; CAA36477.1; ALT_SEQ.
DR   EMBL; M35093; AAB59612.1; ALT_SEQ.
DR   EMBL; Z17324; CAA78972.1; -.
DR   EMBL; Z17325; CAA78973.1; -.
DR   EMBL; M31823; AAA35757.1; -.
DR   EMBL; S81781; AAD14376.1; ALT_INIT.
DR   EMBL; S81736; AAD14369.1; ALT_INIT.
DR   EMBL; M21868; AAA59874.1; ALT_SEQ.
DR   PIR; A35175; A35175.
DR   PIR; B35175; B35175.
DR   PIR; S10218; S10218.
DR   GlycoSuiteDB; P15941; -.
DR   MIM; 158340; -.
DR   MIM; 113720; -.
DR   InterPro; IPR000082; SEA.
DR   Pfam; PF01390; SEA; 1.
DR   SMART; SM00200; SEA; 1.
DR   PROSITE; PS50024; SEA; 1.
KW   Glycoprotein; Signal; Cytoskeleton; Actin-binding; Transmembrane;
KW   Repeat; Alternative splicing.
FT   SIGNAL        1     23       POTENTIAL.
FT   CHAIN        24   1255       MUCIN 1.
FT   DOMAIN       24   1162       EXTRACELLULAR (POTENTIAL).
FT   TRANSMEM   1163   1186       POTENTIAL.
FT   DOMAIN     1187   1255       CYTOPLASMIC (POTENTIAL).
FT   DOMAIN       81    960       44 X 20 AA TANDEM REPEATS.
FT   DOMAIN     1034   1151       SEA.
FT   CARBOHYD    957    957       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    975    975       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1029   1029       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1055   1055       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1133   1133       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   VARSPLIC     19     19       T -> TATTAPKPAT (IN ISOFORM B).
FT   VARSPLIC     20     22       MISSING (IN ISOFORM C).
FT   VARSPLIC     20     31       MISSING (IN ISOFORM D).
FT   VARSPLIC    126    905       MISSING (IN POLYMORPHIC EPITHELIAL
FT                                ISOFORM).
FT   VARSPLIC   1077   1087       FLQIYKQGGFL -> VSIGLSFPMLP (IN SECRETED
FT                                ISOFORM).
FT   VARSPLIC   1088   1255       MISSING (IN SECRETED ISOFORM).
FT   CONFLICT      2      2       T -> A (IN REF. 11).
FT   CONFLICT    134    134       P -> Q (IN REF. 9).
FT   CONFLICT    154    154       P -> Q (IN REF. 9).
FT   CONFLICT   1021   1021       S -> T (IN REF. 3).
FT   CONFLICT   1251   1251       A -> T (IN REF. 3).
SQ   SEQUENCE   1255 AA;  122072 MW;  5E28DFC4DE7D9A82 CRC64;

  Query Match             8.9%;  Score 651;  DB 1;  Length 1255;
  Best Local Similarity   27.8%;  Pred. No. 9.3e-19;
  Matches  295;  Conservative   72;  Mismatches  460;  Indels  234;  Gaps   51;

Qy    236 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 294
          | |||           | :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db     16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy    295 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTI---K 334
          |    :||:|    |    ||               | |     ||||     |:    |
Db     66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVTSAPDNK 125

Qy    335 SAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEP------APTTPKEPAPTTTKEPAPTTTK 388
          || :   ||   | ||| |   ||| :|  |        || |   ||| :|  ||   |
Db    126 PAPGSTAPPAHGVT-SAPDT--RPAPGSTAPPAHGVTSAPDT--RPAPGSTAPPAHGVT- 179

Qy    389 SAPTTPKEPAPTTP----------KKPAPTTPKEP------APTTPKEPTPTTP------ 426
          ||| |   |  | |            :||| :   |      || |   |  | |
Db    180 SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT 239

Qy    427 ----KEPAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPK----E 467
              ||| :  |        || |   |  |||        || |   |  | |
Db    240 SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT 299

Qy    468 PAPTTTKEPSPTTPKEPAPTT---TKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEP-- 522
          || |   |  | |        |:   |: || :|  ||    | ||| |   |:| :|  |
Db    300 SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT-SAPDT--RPAPGSTAPPAH 356

Qy    523 ----APTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEP 570
              || |   |  | |        || |   |  | |        || |   |  |||
Db    357 GVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAH 416

Qy    571 APTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTT 630
            |:  :| |       |      || |   |  | |    |       | |   |  |
Db    417 GVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAP----PAHGVTSAPDTRPAPGSTA 472

Qy    631 PKA----AAPNTPKEPAPTTPK----EPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEP 682
          | |    :||:|   |  | |       || |   |  | |    |     :|| |   |
Db    473 PPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAP----PAHGVTSAPDTRPAP 528

Qy    683 APTTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPT 730
            | |           :|||   ||      ||    :|||  | |     |:    ||
Db    529 GSTAPPAHGVTSAPDTRPAPGSTAP-PAHGVTSAPDTRPAPGSTAPPAHGVTS----APD 583

Qy    731 TPKEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTT 776
          |   | | |      :|| |   |  | |           :|||   ||    | ||
Db    584 TRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAP-PAHGVTSAP 642

Qy    777 SDKPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVS 820
            :|||  | |      :|| |   |  | |           :||| :  | ||     |
Db    643 DTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGS--TAPPAHGVTS 700
```

```
Db    588 TTESSSAPVPT----PSSSTTESSSAPAPTPSSSTTESSSAPVT----SSTTESSSAP-V 638

Qy    599 PEKPAPTTPEELAPT-TPEEPTPTTPEEPAP-----TTPKAAAPNTPKEPAPTTPKEPAP 652

Db    639 PTPSSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESSSAPVT----SSTTESSSAP 694

Qy    653 TTPKEPAPTTPKETAPT-TPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKP 711

Db    695 VT----SSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESSSAPVPTPS-SSTTESSS 749

Qy    712 APTTPKGTAPTTPKEPAP---TTPKEPAPT-TPKGTAPTTLKEPAPTTPKKPAPKELAPT 767

Db    750 APVTSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESSVAPV 809

Qy    768 TTKGPTSTTSDKPAPTTPKETAPTTPKEPAP-----TTPKKPAPTTPETPPPTTSEVSTP 822

Db    810 PTPSSSSNITSSAPSSTPFSSSTESSSVPVPTPSSSTTESSSAPVSSSTTESSVAPVPTP 869

Qy    823 T-----TTKEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTA 877

Db    870 SSSSNITSSAPSSI---PFSSTTESFSTGT-TVTPSSSKYPGSQTETSVSSTTETTIVPT 925

Qy    878 KDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITAT--TTQVTSTTTQDTTPFKI 935

Db    926 KTTTSVTTPSTTTITTTVCSTGTNSAGETTSGCSPKTVTTTVPTTTTTSVTTSSTTTITT 985

Qy    936 TTLKTTTLAPKVTT---TKKTITTT-EIMNKPEETAKPKDRATNSKATTPKPQKP----- 986

Db    986 TVCSTGTNSAGETTSGCSPKTITTTVPCSTSPSETA--------SESTTTSPTTPVTTVV 1037

Qy    987 -----------TKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAM 1034

Db   1038 STTVVTTEYSTSTKPGGEITTTFVTKNIPTTYLTTIAPTP--SVTTVTNFTPTT-ITTTV 1094

Qy   1035 LQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQG 1094

Db   1095 CSTGT------------NSAGETTSGCSPKTVTTTVPCSTGTGEYTTEATTLVTTAVTT 1141

Qy   1095 IIINPMLSDETNICNGKPVDGLTTLRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGIP 1154

Db   1142 TVVTTESSTGTNSA-GKTTTGYTTKSVPTT------YVTTLAPSAPVTPATN-----AVP 1189

Qy   1155 SPIDTVFTRC----NCEGKT 1170

Db   1190 TTITT--TECSAATNAAGET 1207
```

RESULT 4
MUC1_HUMAN
ID MUC1_HUMAN STANDARD; PRT; 1255 AA.
AC P15941; P15942; P13931; P17626; Q14128; Q16442; Q16437; Q9Y4J2;
DT 01-JAN-1990 (Rel. 13, Created)
DT 01-APR-1990 (Rel. 14, Last sequence update)
DT 20-AUG-2001 (Rel. 40, Last annotation update)
DE MUCIN 1 PRECURSOR (POLYMORPHIC EPITHELIAL MUCIN) (PEM) (PEMT)
DE (EPISIALIN) (TUMOR-ASSOCIATED MUCIN) (CARCINOMA-ASSOCIATED MUCIN)
DE (TUMOR-ASSOCIATED EPITHELIAL MEMBRANE ANTIGEN) (EMA) (H23AG) (PEANUT-
DE REACTIVE URINARY MUCIN) (PUM) (BREAST CARCINOMA-ASSOCIATED ANTIGEN
DE DF3).
GN MUC1.
OS Homo sapiens (Human).
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX NCBI_TaxID=9606;
RN [1]
RP SEQUENCE FROM N.A.
RC TISSUE=Pancreas;
RX MEDLINE=90368716; PubMed=2394722;
RA Lan M.S., Batra S.K., Qi W.-N., Metzgar R.S., Hollingsworth M.A.;
RT "Cloning and sequencing of a human pancreatic tumor mucin cDNA.";
RL J. Biol. Chem. 265:15294-15299(1990).
RN [2]
RP SEQUENCE FROM N.A.
RX MEDLINE=90202794; PubMed=2318825;
RA Ligtenberg M.J.L., Vos H.L., Gennissen A.M.C., Hilkens J.;
RT "Episialin, a carcinoma-associated mucin, is generated by a
RT polymorphic gene encoding splice variants with alternative amino
RT termini.";
RL J. Biol. Chem. 265:5573-5578(1990).
RN [3]
RP SEQUENCE FROM N.A.
RC TISSUE=Breast carcinoma;
RX MEDLINE=90368715; PubMed=1697589;
RA Gendler S.J., Lancaster C.A., Taylor-Papadimitriou J., Duhig T.,
RA Peat N., Burchell J., Pemberton L., Lalani E.-N., Wilson D.;
RT "Molecular cloning and expression of human tumor-associated
RT polymorphic epithelial mucin.";
RL J. Biol. Chem. 265:15286-15293(1990).
RN [4]
RP SEQUENCE FROM N.A.
RX MEDLINE=91097524; PubMed=2268309;
RA Lancaster C.A., Peat N., Duhig T., Wilson D.,
RA Taylor-Papadimitriou J., Gendler S.J.;
RT "Structure and expression of the human polymorphic epithelial mucin
RT gene: an expressed VNTR unit.";
RL Biochem. Biophys. Res. Commun. 173:1019-1029(1990).
RN [5]
RP SEQUENCE FROM N.A.
RC TISSUE=Breast carcinoma;
RX MEDLINE=90276413; PubMed=2351132;
RA Wreschener D.H., Hareuveni M., Tsarfaty I., Smorodinsky N.,
RA Horov J., Zaretsky J., Kotkes P., Weiss M., Lathe R., Dion A.,
RA Keydar I.;
RT "Human epithelial tumor antigen cDNA sequences. Differential splicing
RT may generate multiple protein forms.";
RL Eur. J. Biochem. 189:463-473(1990).
RN [6]
RP SEQUENCE FROM N.A.
RC TISSUE=Breast carcinoma;
RX MEDLINE=90276414; PubMed=2112460;
RA Hareuveni M., Tsarfaty I., Zaretsky J., Kotkes P., Horev J.,
RA Zrihan S., Weiss M., Green S., Lathe R., Keydar I., Wreschner D.H.;
RT "A transcribed gene, containing a variable number of tandem repeats,
RT codes for a human epithelial tumor antigen. cDNA cloning, expression
RT of the transfected gene and over-expression in breast cancer
RT tissue.";
RL Eur. J. Biochem. 189:475-486(1990).
RN [7]
RP SEQUENCE FROM N.A.
RX MEDLINE=91033045; PubMed=1688329;
RA Tsarfaty I., Hareuveni M., Horev J., Zaretsky J., Weiss M.,
RA Jeltsch J.M., Garnier J.M., Lathe R., Keydar I., Wreschner D.H.;
RT "Isolation and characterization of an expressed hypervariable gene
RT coding for a breast-cancer-associated antigen.";
RL Gene 93:313-318(1990).
RN [8]
RP PARTIAL SEQUENCE FROM N.A.
RX MEDLINE=88330762; PubMed=3417635;
RA Gendler S.J., Taylor-Papadimitriou J., Duhig T., Rothbard J.,
RA Burchell J.;
RT "A highly immunogenic region of a human polymorphic epithelial mucin
RT expressed by carcinomas is made up of tandem repeats.";
RL J. Biol. Chem. 263:12820-12823(1988).
RN [9]
RP SEQUENCE OF 1-169 FROM N.A.
RX MEDLINE=90088473; PubMed=2597151;
RA Abe M., Siddiqui J., Kufe D.;
RT "Sequence analysis of the 5' region of the human DF3 breast
RT carcinoma-associated antigen gene.";
RL Biochem. Biophys. Res. Commun. 165:644-649(1989).
RN [10]
RP SEQUENCE OF 1-109 FROM N.A.
RC TISSUE=Thyroid;
RX MEDLINE=96183746; PubMed=8608966;
RA Weiss M., Baruch A., Keydar I., Wreschner D.H.;

```
Qy     796 PAPTTPKKPAPT-TPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENS 854
           | |:   :| |: ||| | || :     || : |||      || ||  | ||||
Db    1254 PTPS--DEPTPSETPEEPIPTDTPSDEPTPSDEPT----PSDEPTP--SDEPTPSDEPTP 1305

Qy     855 PKEPGVPTTKTPAAT--KPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTE 912
             ||      ||: |  :|  | |  |: | |      | :  |  : |   : | |
Db    1306 SDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSD----EPTPSDEPTPSDEPTPSDEPTPS 1361

Qy     913 SKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDR 972
           : | :     | | :: ||       |||  |  :||                      |
Db    1362 DEPTPSDEPTPSETPEEPTP-------TTTPTPTPSTT-----------------PTSG 1396

Qy     973 ATNSKAT--------TPKPQKPTKAP-KKPTSTKKPKTMPRVRKPKTTPTPRKM------ 1017
           : |  :           |  ||  |  ||||| |     : :|  :  | :
Db    1397 SGGSGGSGGGGGGGGGTVPTSPTPTPTSKPTSTPAP---TEIEEPTPSDVPGAIGGEHRA 1453

Qy    1018 -------TSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHM- 1069
                |  || | |   |  :           :|| :     :::: ||:  :|:
Db    1454 YLRGYPDGSFRPERNITRAEAAVIF----------AKLL----GADESYGAQSASPYSD 1498

Qy    1070 LLRPH--------------------VFMPE--------VTPDMDYLPRVPNQGI----- 1095
           |  |                      | |:        |  : :| :|  | |
Db    1499 LADTHWAAWAIKFATSQGLFKGYPDGTFKPDQNITRAEFATVVLHFLTKVKGQEIMSKLA 1558

Qy    1096 ---IINPMLSDETNICNG----KPVDGLTTL------RNGTLVAFRGHYFWMLSPFSPPS 1142
             | ||   |    | |    : :: ||:|       :||              | | :
Db    1559 TIDISNPKFDD----CVGHWAQEFIEKLTSLGYISGYPDGT-------------FKPQN 1600

Qy    1143 PARRITEVWGIPSPIDTVFTRCNCEGKTFFFKD--SQYWRFTNDIKD 1187
             :|  |    : |:    |    |    | |    || |  || |
Db    1601 YIKRSESV----ALINRALERGPLNGAPKLFPDVNESYWAF-GDIMD 1642
```

```
RESULT   3
AMYH_YEAST
ID   AMYH_YEAST     STANDARD;      PRT;  1367 AA.
AC   P08640; P08068;
DT   01-AUG-1988 (Rel. 08, Created)
DT   01-FEB-1995 (Rel. 31, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   GLUCOAMYLASE S1/S2 PRECURSOR (EC 3.2.1.3) (GLUCAN 1,4-ALPHA-
DE   GLUCOSIDASE) (1,4-ALPHA-D-GLUCAN GLUCOHYDROLASE).
GN   STA1 OR STA2 OR MAL5 OR YIR019C.
OS   Saccharomyces cerevisiae (Baker's yeast).
OC   Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
OC   Saccharomycetales; Saccharomycetaceae; Saccharomyces.
OX   NCBI_TaxID=4932;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=S288C / AB972;
RA   Barrell B.G., Badcock K., Bankier A.T., Bowman S., Brown D.,
RA   Churcher C.M., Connor R., Copsey T., Dear S., Devlin K., Fraser A.,
RA   Gentles S., Hamlyn N., Horsnell T.S., Hunt S., Jagels K., Jones M.,
RA   Louis E., Lye G., Moule S., Moule T., Odell C., Pearson D.,
RA   Rajandream M.A., Riles L., Rowley N., Skelton J., Smith V.,
RA   Walsh S.V., Whitehead S.;
RL   Submitted (DEC-1994) to the EMBL/GenBank/DDBJ databases.
RN   [2]
RP   SEQUENCE OF 1-242 AND 762-1331 FROM N.A.
RX   MEDLINE=87194600; PubMed=3106330;
RA   Yamashita I., Nakamura M., Fukui S.;
RT   "Gene fusion is a possible mechanism underlying the evolution of
RT   STA1.";
RL   J. Bacteriol. 169:2142-2149(1987).
RN   [3]
RP   SEQUENCE OF 1-31 FROM N.A.
RC   STRAIN=SPX101-1C;
RX   MEDLINE=89031230; PubMed=3141213;
RA   Pardo J.M., Ianez E., Zalacain M., Claros M.G., Jimenez A.;
RT   "Similar short elements in the 5' regions of the STA2 and SGA genes
RT   from Saccharomyces cerevisiae.";
RL   FEBS Lett. 239:179-184(1988).
```

```
CC   -!- CATALYTIC ACTIVITY: HYDROLYSIS OF TERMINAL 1,4-LINKED ALPHA-D-
CC       GLUCOSE RESIDUES SUCCESSIVELY FROM NON-REDUCING ENDS OF THE CHAINS
CC       WITH RELEASE OF BETA-D-GLUCOSE.
CC   -!- SIMILARITY: TO S.POMBE SPBC215.13.
CC   -!- SIMILARITY: SOME, TO S.POMBE SPCC285.13C.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; Z38061; CAA86176.1; -.
DR   EMBL; M16164; AAA35014.1; -.
DR   EMBL; M16165; AAA35015.1; -.
DR   EMBL; X13857; CAA32069.1; -.
DR   PIR; B26877; B26877.
DR   PIR; A26877; A26877.
DR   PIR; S48478; S48478.
DR   SGD; S0001458; MUC1.
KW   Hydrolase; Glycosidase; Polysaccharide degradation; Glycoprotein;
KW   Signal; Multigene family.
FT   SIGNAL       1     21       POTENTIAL.
FT   CHAIN       22   1367       GLUCOAMYLASE S1/S2.
FT   DOMAIN     210   1367       SER/THR-RICH.
FT   CARBOHYD   817    817       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   874    874       N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE   1367 AA;  136110 MW;  91C00E2DBD61AA9D CRC64;
```

```
  Query Match             11.1%;  Score 810.5;  DB 1;  Length 1367;
  Best Local Similarity   27.2%;  Pred. No. 7.4e-25;
  Matches 332;  Conservative  114;  Mismatches  519;  Indels  255;  Gaps   54;

Qy      23 SSQDLSS-CAGRCGEGYSRDA------TCNCDYNCQHYMECCPDFKRVCTAAEVHNPTSP 75
           :| |||: |     :|:|:          :|| ||
Db     171 TSFDLSTGCNNYDNQGHSQTDFPGFYWNIDCDNNC---------------------- 205

Qy      76 PSSKKAPPPSGASQTIKSTTKRSPKPPNKKKTKKVIESEEITEEHSVSENQESSSSSSSS 135
                   | :   ||:|  |    :   |    ||   |   | |    ||:| ||:
Db     206 ----------GGT---KSSTTTSSTSESSTTTSSTSESSTTTSSTSESSTTTSSTSESST 252

Qy     136 SSSSTIWKIKSSKNSAANRELQKKLKVKDNKKNRTKKKPTPKPPVVDEAGSGLDNGDFKV 195
           |||:|      :  :       |:        : ||:|||  ||  |
Db     253 SSSTT------APATPTTTSCTKEKPTPPTTTSCTKEKPT--PPHHD------------- 291

Qy     196 TTPDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTN 255
           |||  |      | :|: |  | |    | |:  |:| |:: :|       | |  ::||
Db     292 TTPCT-----KKKTTTSKTCTKKTTTPVPT--PSSSTTESSS------APVPTPSSSTT- 337

Qy     256 KQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPT 315
                ::::||  | |  ::|:   ||       ||| : ||    |  |      :
Db     338 ---------ESSSAPVTSSTTESSS---APV------PTPSSSTTESSSAPVTSSTTESS 379

Qy     316 TPKEPASTTPKEPTPTTIKSAPTTPKEPAPTT-----TKSAP----TTPKEPAPTTTKEP 366
           :    :|||      |    |: ||    || |     : |||    ||    || |:
Db     380 SAPVTSSTTESSSAPVPTPSSSTTESSSAPVTSSTTESSSAPVTSSTTESSSAPVTSS-- 437

Qy     367 APTTPKEPAP----TTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPT 422
             ||    ||    ||    ||  | |: ||    || |     : ||    ||    ||
Db     438 --TTESSSAPVTSSTTESSSAPVPTPSSSTTESSSAPVT----SSTTESSSAPV----PT 487

Qy     423 P---TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPSPT 479
           |   ||     || |   : ||    ||          ||       :   |||| :   | |
Db     488 PSSSTTESSSAPVT---SSTTESSSAPV-------PTPSSSTTESSSAPAPTPS---SST 534

Qy     480 TPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPT-TPKKP 538
           |    || |:    ||    ||  | |: ||    :| |:      ||    ||  ||
Db     535 TESSSAPVTSS---TTESSSAPVPTPSSSTTESSSTPVTSS----TTESSSAPVPTPSSS 587

Qy     539 APTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTT 598
```

```
Qy     509 TTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPK 568
           |||  |  |||  |: ||   | |||   |:|||   |  ||     ||||  |  | |
Db    1703 TTPSSPI-TTTTTPSSTTTPSPPPTTMTTPSPTTTPSPPTTTMTTLPPTTTSSPLTTTPL 1761

Qy     569 EPAPTTP-----KETAPTTP---------------------------------------- 583
            |: | |        | ||||
Db    1762 PPSITPPTFSPFSTTTPTTPCVPLCNWTGWLDSGKPNFHKPGGDTELIGDVCGPGWAANI 1821

Qy     584 ---KKLTP-------------------------------------------------TT 590
               : |                                                    |
Db    1822 SCRATMYPDVPIGQLGQTVVCDVSVGLICKNEDQKPGGVIPMAFCLNYEINVQCCECVTQ 1881

Qy     591 PEKLAPTTPEKPAPTTPEELAPTTPEEPTPT-TPEEPAPTTPKAAAPNTPKEPAPTTPKE 649
           |  :  || | | | |   :  ||   |||| |  :  |||          | ||  :
Db    1882 PTTMTTTTTENPTPPTTTPITTTTTTVTPTPTPTGTQTPTTTPITTTTTTVTPTPTPTGTQT 1941

Qy     650 P-----APTTPKEPAPT-TPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEP 703
           |       ||   | || | :|  |||  |  |    | ||  : |    :  |||  |
Db    1942 PTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTTVTPTPTPTGTQTPTTTPITTTTTVTP 2001

Qy     704 TSTTSDKPAP-TTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPK 762
           | | :    | |||  |  |    | ||  : |  |||  |  |    | ||  : |
Db    2002 TPTPTGTQTPTTTPITTTTTTVTPTPTPTGTQTPT-TTPITTTTTTVTPTPTPTGTQTPTTT 2060

Qy     763 ELAPTTTKGPTSTTSDKPAP-TTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVST 821
             :  |||  || | :    | |||  |  |    | ||  : |  |||  |  |  :  |
Db    2061 PITTTTTTVTPTPTPTGTQTPTTTPITTTTTTVTPTPTPTGTQTPT-TTPITTTTTTVTPTPT 2119

Qy     822 PTTTKEPTTIHKSPDESTPELSAEPTPKALENSPKEP----------GVPT-TKTPAATK 870
           || |: |||   :|  :|  ::  |||   :       |           || |:||  |
Db    2120 PTGTQTPTT---TPITTTTTTVTPTPTPTGTQTPTTTPITTTTTTVTPTPTPTGTQTPTTTP 2176

Qy     871 PEMTTTAKDKTTERDLR---TTP--ETTTAAPKMT---KETATTTEKTTESKITATTTQV 922
             |||      |    :  |||   |||  |  |    :| |||  || : :| | | |
Db    2177 ITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTP- 2235

Qy     923 TSTTTQDTTPFKITTLKTTTLAPKVTTT-KKTITTTEIMNKPEETAKPKDRATNSKATTP 981
           | | |  |||  |||  |||: |  | |  :| ||| |      |  |    | :  |||
Db    2236 TGTQTPTTTP--ITT--TTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTP 2291

Qy     982 KPQKPTKAP-KKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNP---TSRIAEAMLQT 1037
             |  |   || |: | | |       ||||    :  |   |   |: :       |
Db    2292 ITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPT 2351

Qy    1038 TTR-PNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTP 1082
            |: |  || :    | |     |    :||     |      |||
Db    2352 GTQTPTTTPITTTTTVTPTPTPTG---TQTP--TTTPITTTTTVTP 2392
```

```
RESULT   2
SLP1_CLOTM
ID   SLP1_CLOTM     STANDARD;      PRT;  1664 AA.
AC   Q06852;
DT   01-JUN-1994 (Rel. 29, Created)
DT   01-JUN-1994 (Rel. 29, Last sequence update)
DT   01-FEB-1996 (Rel. 33, Last annotation update)
DE   CELL SURFACE GLYCOPROTEIN 1 PRECURSOR (OUTER LAYER PROTEIN B) (S-LAYER
DE   PROTEIN 1).
GN   OLPB.
OS   Clostridium thermocellum.
OC   Bacteria; Firmicutes; Bacillus/Clostridium group; Clostridiaceae;
OC   Clostridium.
OX   NCBI_TaxID=1515;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=NCIB 10682;
RX   MEDLINE=93209931; PubMed=8458832;
RA   Fujino T., Beguin P., Aubert J.-P.;
RT   "Organization of a Clostridium thermocellum gene cluster encoding the
RT   cellulosomal scaffolding protein CipA and a protein possibly involved
RT   in attachment of the cellulosome to the cell surface.";
RL   J. Bacteriol. 175:1891-1899(1993).
```

```
CC   -!- SUBUNIT: ASSEMBLED INTO MONO-LAYERED CRYSTALLINE ARRAYS.
CC   -!- SUBCELLULAR LOCATION: CELL WALL.
CC   -!- SIMILARITY: CONTAINS 4 S-LAYER HOMOLOGY (SLH) DOMAINS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X67506; CAA47841.1; -.
DR   InterPro; IPR001119; SLH.
DR   Pfam; PF00395; SLH; 3.
DR   PROSITE; PS01072; SLH_DOMAIN; 2.
KW   Cell wall; S-layer; Signal; Repeat.
FT   SIGNAL        1     28       POTENTIAL.
FT   CHAIN        29   1664       CELL SURFACE GLYCOPROTEIN 1.
FT   DOMAIN       36    763       4 X 156 AA APPROXIMATE REPEATS.
FT   REPEAT       36    191       1.
FT   REPEAT      207    363       2.
FT   REPEAT      409    565       3.
FT   REPEAT      607    763       4.
FT   DOMAIN      771   1377       APPROXIMATE TANDEM REPEATS OF
FT                                T-P-S-D-E-P.
FT   DOMAIN     1378   1449       GLY/PRO/SER/THR-RICH.
FT   DOMAIN     1453   1494       SLH 1 (INCOMPLETE).
FT   DOMAIN     1495   1565       SLH 2.
FT   DOMAIN     1566   1625       SLH 3.
FT   DOMAIN     1626   1646       SLH 4 (INCOMPLETE).
SQ   SEQUENCE   1664 AA;  178194 MW;  5F396695BA9FE74B CRC64;
```

```
  Query Match           13.0%;  Score 950;  DB 1;  Length 1664;
  Best Local Similarity 31.0%;  Pred. No. 3.8e-30;
  Matches  312;  Conservative  104;  Mismatches  361;  Indels  230;  Gaps   55;

Qy     289 VLAKPTP-KAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTT 347
           |: :| | || :    |  |   ||||:          ||||:    ||  || |:
Db     758 VVIQPAPIKAASDEPIPTDTPSDEPTPS----------DEPTPS---DEPTPSDEPTPSD 804

Qy     348 TKSAPTTPKEPAPTTTKEPAPTTPKEPA----PTTTKEPAPTTTKSAPTTPKEPAPTTPK 403
              :  ||:|| || |    ||  ||      || : || |:   :  ||:|| ||
Db     805 EPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTP 864

Qy     404 KPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPT-TPKEPAPTAPKKPAPTTPKEPAP 462
             ||   || |:  ||||:   || | : || |: ||:|| ||      ||   || |
Db     865 SDEPTPSDEPTPS--DEPTPS--DEPTP-SDEPTPSETPEEPIPTDTPSDEPTPSDEPTP 919

Qy     463 TTPKEPA--PTTTKEPSPT-TPKEPAPTTTKS-APTTTKEPAPT---TTKSAPTTPKEPS 515
           :    |:  || : ||:|: ||:|| || | |  || : || |:   |    ||  ||:
Db     920 SDEPTPSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPT 979

Qy     516 ----PTTTKEPAPT-TPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEP 570
               || : || |: ||:|| ||      ||   || |:   |:  |   ||  ||
Db     980 PSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEP 1039

Qy     571 APT---TPKETAPTTPKKLTPTTPEKLAPTTPEKPAPT---TPEELAPTTPEEPTPT-TP 623
            |:   || |    ||::  ||      ||  ::| |:   || :  ||  :||||: ||
Db    1040 TPSDEPTPSE----TPEEPIPTDTPSDEPTPSDEPTPSD-EPTPSDEPTPSETP 1094

Qy     624 EEPAPTTPKAAAPNTPKEPAPTTPKEPA--PTTPKEPAPT-TPKETAPTTPKGTAPTTLK 680
           ||| ||   :  |    || |:    |:  ||   || |: ||:|  ||      ||
Db    1095 EEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSD 1154

Qy     681 EPAPTTPKKPAPKELAPTTTKEPT-STTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTT 739
           || |:   :| | :  || : ||| | | ::| ||      ||   || |:   || |:
Db    1155 EPTPS--DEPTPSD-EPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPS--DEPTPS- 1208

Qy     740 PKGTAPTTLKEPAPT-TPKKPAPKELAPTTTKGPTSTTSDKPAPT---TPKETAPTTPKE 795
             ||   || |: ||::|           || | ||:| |:   ||  :  ||   |
Db    1209 ---DEPTPSDEPTPSETPEEPI----------PTDTPSDEPTPSDEPTPSD-EPTPSDE 1253
```

```
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------------
DR   EMBL; L21998; AAB95295.1; -.
DR   EMBL; M74027; AAA59875.1; -.
DR   EMBL; M94131; AAA59163.1; -.
DR   EMBL; M94132; AAA59164.1; -.
DR   MIM; 158370; -.
DR   InterPro; IPR000359; Cys_knot.
DR   InterPro; IPR000561; EGF-like.
DR   InterPro; IPR002400; GF_cysknot.
DR   InterPro; IPR001007; VWFC.
DR   InterPro; IPR001846; Vwd.
DR   Pfam; PF00007; Cys_knot; 1.
DR   Pfam; PF00094; vwd; 4.
DR   PRINTS; PR00438; GFCYSKNOT.
DR   SMART; SM00214; VWC; 2.
DR   SMART; SM00011; VWC_def; 2.
DR   SMART; SM00216; VWD; 4.
DR   PROSITE; PS00022; EGF_1; UNKNOWN_1.
DR   PROSITE; PS01185; CTCK_1; 1.
DR   PROSITE; PS01225; CTCK_2; 1.
DR   PROSITE; PS01208; VWFC; 2.
KW   Glycoprotein; Repeat; Signal.
FT   SIGNAL        1     20       POTENTIAL.
FT   CHAIN        21   5179       MUCIN 2.
FT   DOMAIN     1401   1747       APPROXIMATE REPEATS.
FT   REPEAT     1401   1416       1.
FT   REPEAT     1417   1432       2.
FT   REPEAT     1433   1448       3.
FT   REPEAT     1449   1464       4.
FT   REPEAT     1465   1471       5.
FT   REPEAT     1472   1478       6.
FT   REPEAT     1479   1494       7A.
FT   REPEAT     1495   1517       7B.
FT   REPEAT     1518   1533       8A.
FT   REPEAT     1534   1556       8B.
FT   REPEAT     1557   1572       9A.
FT   REPEAT     1573   1596       9B.
FT   REPEAT     1597   1612       10A.
FT   REPEAT     1613   1635       10B.
FT   REPEAT     1636   1651       11A.
FT   REPEAT     1652   1675       11B.
FT   REPEAT     1676   1683       12.
FT   REPEAT     1684   1699       13.
FT   REPEAT     1700   1715       14.
FT   REPEAT     1716   1731       15.
FT   REPEAT     1732   1747       16.
FT   DOMAIN     4815   4886       VWFC 1.
FT   DOMAIN     4924   4991       VWFC 2.
FT   DOMAIN     5075   5160       CTCK.
FT   DISULFID   5075   5122       BY SIMILARITY.
FT   DISULFID   5089   5136       BY SIMILARITY.
FT   DISULFID   5098   5152       BY SIMILARITY.
FT   DISULFID   5102   5154       BY SIMILARITY.
FT   DISULFID      ?   5159       BY SIMILARITY.
FT   CARBOHYD    163    163       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    423    423       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    670    670       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    770    770       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    894    894       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1139   1139       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1154   1154       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1215   1215       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1230   1230       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1246   1246       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1787   1787       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1820   1820       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4339   4339       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4351   4351       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4362   4362       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4373   4373       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4422   4422       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4438   4438       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4502   4502       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4616   4616       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4627   4627       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4752   4752       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4787   4787       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4881   4881       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4888   4888       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4955   4955       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4970   4970       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   5019   5019       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   5038   5038       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   5069   5069       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CONFLICT   1351   1351       H -> L (IN REF. 3).
FT   CONFLICT   1412   1412       T -> S (IN REF. 3).
FT   CONFLICT   1449   1449       L -> P (IN REF. 3).
FT   CONFLICT   1504   1504       M -> T (IN REF. 3).
FT   CONFLICT   4192   4192       G -> S (IN REF. 2).
SQ   SEQUENCE   5179 AA;  540295 MW;  85CD7571FB9A5663 CRC64;

  Query Match             16.1%;  Score 1171;  DB 1;  Length 5179;
  Best Local Similarity   29.2%;  Pred. No. 3.1e-38;
  Matches  399;  Conservative   79;  Mismatches  524;  Indels  364;  Gaps   52;

Qy      29 SCAGRCGEGYSRDATCNCDYNCQHYMEC--------------CPDFKRVCTAAEVHNPTS 74
           ||:  |  |   |  | |     :  ||              || |       : :||
Db    1079 SCS--CDTG--GDCECFCSAVASYAQECTKEGACVFWRTPDLCPIF------CDYYNPPH 1128

Qy      75 PPSSKKAPPPSGASQTIKSTT---------------KRSPK--PPNKKKTKKVIESEE-- 115
             |  : : :| ::                     | ||  |  ::  || : :::
Db    1129 ECEWHYEPCGNRSFETCRTINGIHSNISVSYLEGCYPRCPKDRPIYEEDLKKCVTADKCG 1188

Qy     116 --ITEEH----SVSENQESSSSSSSSSSSTI------------------WKIKSSKNS 150
             : : |    :    :|:  |   ::||  :                  |:|
Db    1189 CYVEDTHYPPGASVPTEETCKSCVCTNSSQVVCRPEEGKILNQTQDGAFCYWEI-----C 1243

Qy     151 AANRELQK-------------------KLKVKDNKKNRTKKKPTPKPPVV--------- 181
             |  ::|                    |       |    ||   |
Db    1244 GPNGTVEKHFNICSITTRPSTLTTFTTITLPTTPTSFTTTTTTTPTSSTVLSTTPKLCC 1303

Qy     182 --------DEAGSGLDNGDFK----------------VTTPDTSTTQH-NKVSTSPK--- 213
                 |   || |:|| :                 |  |  |  ||  ||
Db    1304 LWSDWINEDHPSSGSDDGDREPFDGVCGAPEDIECRSVKDPHLSLEQHGQKVQCDVSVGF 1363

Qy     214 -------------------------------ITTAKPINPRPSLPPNSDTSKETSLTVN 241
                                           |||  |    || || : |:   : | :
Db    1364 ICKNEDQFGNGPFGLCYDYKIRVNCCWPMDKCITTPSPPTTTPSPPPTTTTTLPPTTTPS 1423

Qy     242 KETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTT 301
             ||     ||||  |:|     ||:                ||:     |:|   |||
Db    1424 PPTT-----TTTTPPPTTTPSPPITTTTTP-----------LPTT----TPSPPISTTT 1462

Qy     302 KGPALTTPKEPT-----PTTPKEPASTTPKEPTPTTIKSAP-TTPKEP-------APTTT 348
             |  |||  ||      |||   | :||   | |||  | | |||  |          ||||
Db    1463 TPPPTTTPSPPTTTPSPPTTTPSPPTTTTTTPPPTTTPSPPMTTPITPPASTTTLPPTTT 1522

Qy     349 KSAP-TTPKEPAPTTTKEPAPTTPKEP-------APTTTKEPAPTTTKSAP--TTPKEPA 398
             | | ||   | ||||  |  |||  |          ||||  | ||||  : |  |||  |
Db    1523 PSPPTTTTTTPPPTTTPSPPTTTPITPPTSTTTLPPTTTPSPPPTTTTTPPPTTTPSPPT 1582

Qy     399 PTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEP-------APTAPKK 451
             |||  |  ||   | |||  || ||   |  ||  |  |||  |          ||
Db    1583 TTTPSPPTITTTTPPPTTTPSPPTTTTTTPPPTTTPSPPTTTPITPPTSTTTLPPTTTPS 1642

Qy     452 PAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAP--- 508
           | |||   | |||   |  |||  |  ||   | |||| |:| ||    |||| : |
Db    1643 PPPTTTTTPPPTTTPSPPTTTTPSPPITTTTPPPTTTPSSPITTTPSPPTTTMTTPSPT 1702
```

GenCore version 4.5
Copyright (c) 1993 - 2000 Compugen Ltd.

OM protein - protein search, using sw model

Run on: April 26, 2002, 16:32:37 ; Search time 62.75 Seconds
(without alignments)
798.154 Million cell updates/sec

Title: AA6
Perfect score: 7294
Sequence: 1 MAWKTLPIYLLLLLSVFVIQ.....ARAITTRSGQTLSKVWYNCP 1366

Scoring table: BLOSUM62
Gapop 10.0 , Gapext 0.5

Searched: 100059 seqs, 36664827 residues

Total number of hits satisfying chosen parameters: 100059

Minimum DB seq length: 0
Maximum DB seq length: 2000000000

Post-processing: Minimum Match 0%
Maximum Match 100%
Listing first 45 summaries

Database : SwissProt_39:*

Pred. No. is the number of results predicted by chance to have a score greater than or equal to the score of the result being printed, and is derived by analysis of the total score distribution.

SUMMARIES

Result No.	Score	% Query Match	Length	DB	ID	Description
1	1171	16.1	5179	1	MUC2_HUMAN	Q02817 homo sapien
2	950	13.0	1664	1	SLP1_CLOTM	Q06852 clostridium
3	810.5	11.1	1367	1	AMYH_YEAST	P08640 saccharomyc
4	651	8.9	1255	1	MUC1_HUMAN	P15941 h mucin 1 p
5	559	7.7	875	1	FP1_MYTED	Q25460 mytilus edu
6	555.5	7.6	2700	1	ZAN_HUMAN	Q9y493 homo sapien
7	551	7.6	620	1	EXTN_TOBAC	P13983 nicotiana t
8	551	7.6	1087	1	NFH_MOUSE	P19246 mus musculu
9	530.5	7.3	1162	1	TCNA_TRYCR	P23253 trypanosoma
10	530	7.3	865	1	CPN_DROME	Q02910 drosophila
11	518.5	7.1	872	1	FP1_MYTCO	Q25434 mytilus cor
12	517.5	7.1	831	1	NFH_RAT	P16884 rattus norv
13	509.5	7.0	662	1	MUC1_XENLA	Q05049 xenopus lae
14	503.5	6.9	1970	1	RPB1_HUMAN	P24928 homo sapien
15	502	6.9	1020	1	NFH_HUMAN	P12036 homo sapien
16	498.5	6.8	1970	1	RPB1_MOUSE	P08775 mus musculu
17	493.5	6.8	467	1	RPB1_CRIGR	P11414 cricetulus
18	490	6.7	2142	1	BAT2_HUMAN	P48634 homo sapien
19	488.5	6.7	826	1	SSP2_PLAYO	Q01443 plasmodium
20	475.5	6.5	267	1	EXTN_MAIZE	P14918 zea mays (m
21	471.5	6.5	5376	1	ZAN_MOUSE	O88799 mus musculu
22	468.5	6.4	634	1	HWP1_CANAL	P46593 candida alb
23	467	6.4	817	1	VRP1_YEAST	P37370 saccharomyc
24	454	6.2	797	1	VGLX_HSVEB	P28968 equine herp
25	454	6.2	1161	1	YJ9P_YEAST	P47179 saccharomyc
26	453.5	6.2	1229	1	N121_HUMAN	Q9y2n3 homo sapien
27	448.5	6.1	670	1	VG50_HSVI1	Q00130 ictalurid h
28	444.5	6.1	2774	1	MAPA_RAT	P34926 rattus norv
29	443.5	6.1	1794	1	YAV1_SCHPO	Q10172 schizosacch
30	442.5	6.1	751	1	FP1_MYTGA	Q27409 mytilus gal
31	439.5	6.0	1083	1	T2D3_HUMAN	O00268 homo sapien
32	436.5	6.0	1411	1	TCOF_HUMAN	Q13428 homo sapien
33	435.5	6.0	3256	1	KI67_HUMAN	P46013 homo sapien
34	432.5	5.9	439	1	XP2_XENLA	P17437 xenopus lae
35	432	5.9	3164	1	TEGU_HSV11	P10220 herpes simp
36	428	5.9	2517	1	NCR2_HUMAN	Q9y618 h nuclear r
37	427.5	5.9	1251	1	YQU3_CAEEL	Q09550 caenorhabdi
38	426	5.8	2715	1	TRX2_HUMAN	Q9umn6 homo sapien
39	425.5	5.8	1185	1	DRPL_HUMAN	P54259 homo sapien
40	424.5	5.8	1125	1	MAP4_MOUSE	P27546 mus musculu
41	424	5.8	307	1	SGS3_DROME	P02840 drosophila
42	423.5	5.8	3421	1	TEGU_HSVEB	P28955 equine herp
43	422.5	5.8	1183	1	DRPL_RAT	P54258 rattus norv
44	419.5	5.8	2476	1	ZAN_PIG	Q28983 sus scrofa
45	411.5	5.6	907	1	VGP3_EBV	P03200 epstein-bar

ALIGNMENTS

```
RESULT   1
MUC2_HUMAN
ID   MUC2_HUMAN      STANDARD;      PRT;  5179 AA.
AC   Q02817; Q14878;
DT   01-JUN-1994 (Rel. 29, Created)
DT   01-NOV-1997 (Rel. 35, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   MUCIN 2 PRECURSOR (INTESTINAL MUCIN 2).
GN   MUC2 OR SMUC.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Intestine;
RX   MEDLINE=94132002; PubMed=8300571;
RA   Gum J.R. Jr., Hicks J.W., Toribara N.W., Siddiki B., Kim Y.S.;
RT   "Molecular cloning of human intestinal mucin (MUC2) cDNA.
RT   Identification of the amino terminus and overall sequence similarity
RT   to prepro-von Willebrand factor.";
RL   J. Biol. Chem. 269:2440-2446(1994).
RN   [2]
RP   SEQUENCE OF 626-1895 AND 4196-5179 FROM N.A.
RC   TISSUE=Colon;
RX   MEDLINE=93016075; PubMed=1400449;
RA   Gum J.R. Jr., Hicks J.W., Toribara N.W., Rothe E.-M., Lagace R.E.,
RA   Kim Y.S.;
RT   "The human MUC2 intestinal mucin has cysteine-rich subdomains located
RT   both upstream and downstream of its central repetitive region.";
RL   J. Biol. Chem. 267:21375-21383(1992).
RN   [3]
RP   SEQUENCE OF 1343-1895 AND 4176-4195 FROM N.A.
RX   MEDLINE=91358717; PubMed=1885763;
RA   Toribara N.W., Gum J.R. Jr., Culhane P.J., Lagace R.E., Hicks J.W.,
RA   Petersen G.M., Kim Y.S.;
RT   "MUC-2 human small intestinal mucin gene structure. Repeated arrays
RT   and polymorphism.";
RL   J. Clin. Invest. 88:1005-1013(1991).
CC   -!- FUNCTION: COATS THE EPITHELIA OF THE INTESTINES, AIRWAYS, AND
CC       OTHER MUCUS MEMBRANE-CONTAINING ORGANS. THOUGHT TO PROVIDE A
CC       PROTECTIVE, LUBRICATING BARRIER AGAINST PARTICLES AND INFECTIOUS
CC       AGENTS AT MUCOSAL SURFACES.
CC   -!- SUBUNIT: MULTIMERIC.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- TISSUE SPECIFICITY: COLON, SMALL INTESTINE, COLONIC TUMORS,
CC       BRONCHUS, CERVIX AND GALL BLADDER.
CC   -!- PTM: ALL CYSTEINE RESIDUES ARE INVOLVED IN INTRACHAIN OR
CC       INTERCHAIN DISULFIDE BONDS (BY SIMILARITY).
CC   -!- POLYMORPHISM: THE NUMBER OF REPEATS IS HIGHLY POLYMORPHIC AND
CC       VARIES AMONG DIFFERENT ALLELES.
CC   -!- SIMILARITY: THE N-TERMINAL DOMAIN SHOWS SOME SIMILARITY TO THAT
CC       OF SILKWORM HEMOCYTIN.
CC   -!- SIMILARITY: CONTAINS 2 VWFC DOMAINS.
CC   -!- SIMILARITY: CONTAINS 1 C-TERMINAL CYSTINE KNOT-LIKE DOMAIN (CTCK).
CC   ---------------------------------------------------------------------
```

A;Residues: 1-489 <ARS>
A;Cross-references: EMBL:X91836; NID:g1015936; PIDN:CAA62943.1; PID:g1015937
A;Experimental source: sub_species Red caloona
R;Arsenijevic-Maksimovic, I.; Broughton, W.J.; Krause, A.
submitted to the EMBL Data Library, April 1995
A;Description: A class of root-hair specific extensins involved in rhizobium/legume inte
A;Reference number: S54155
A;Accession: S54155
A;Status: preliminary
A;Molecule type: mRNA
A;Residues: 326-489 <AR2>
A;Cross-references: EMBL:X86030; NID:g791149; PID:g791150
C;Genetics:
A;Gene: Ext26G
C;Superfamily: hydroxyproline-rich glycoprotein
C;Keywords: glycoprotein; hydroxyproline
F;1-23/Domain: signal sequence #status predicted <SIG>
F;24-489/Product: extensin class 1 #status predicted <MAT>

Query Match 8.7%; Score 633; DB 2; Length 489;
Best Local Similarity 31.2%; Pred. No. 2.8e-20;
Matches 149; Conservative 33; Mismatches 252; Indels 44; Gaps 7;

```
Qy    340 PKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAP 399
          ||:  |    : |   | | |     |:|: |          | |   |  |   | | |
Db     39 PKQTPPYYYNAPPYYYKSPPP-----PSPSPP----------PPPYVHKYPPYYYKSPPP 83

Qy    400 TTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTP---KEPAPTAPKKPAPTT 456
           :|  | |   | | | :|  | |   | | | :  |:|  |   | | | :|  | |
Db     84 PSPSPPPPYVYKSPPPPSPSPPPPYYYKSPPPPS--PSPPPPYYYKSPPPPSPSPPPPYY 141

Qy    457 PKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSP 516
           | | | :|  | |   | | | :|  | |   || |              | :|  | |
Db    142 YKSPPPPSPSPPPPYVYKSPPPPSPSPPPPYYYKSPP-------------PPSPSPPPP 187

Qy    517 TTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPK 576
             | | | :|  | |   | | | :|  | |   | | | :   | |   | | | :|
Db    188 YYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYVYKSPPPPSPSPPPPYYYKSPPPPSPS 247

Qy    577 ETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAP 636
            |   |  | :|    |   : | | :|    |   : | | :|  | |   |:  |
Db    248 PPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPP 307

Qy    637 NTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELA 696
           :|  | |   | | | :|  | |   |   | :|    |   | | | :|  | |
Db    308 PSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYVYKSPPPPSPSPPPPYYY- 366

Qy    697 PTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTP 756
              | |   :   | |   |   | :|  | |   | | | :|    |   | | | :|
Db    367 ----KSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSP 422

Qy    757 KKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETPPP 814
            |       |   | |   :   | |   |   | :|  | |   | | | :|  |||
Db    423 SPP-----PPYVYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPP 475
```

Search completed: April 26, 2002, 16:32:20
Job time: 735 sec

```
Db   2417 SKQTISSTPTPDTTTASEEPTKSTSMSPDLSTTSNVLSESSTTPESSS----KSPVSSST 2472

Qy    647 ----------PKEPAPT---------TPKEPAP-----TTPKETAPTTPKGTAPTTLKE 681
                   | |  |             |  |:|     ||  ||:    : :   :::
Db   2473 EGISVVTSTEFSKVPESTISSVLEEDLTKTTPSPILEETTTASETSEPLTEDSLTVSVRI 2532

Qy    682 PAPTTPKKPAPKELAPTTTKEPTSTTSDKPA---------PT----------------T 715
          ||  :  |||   |||   :|  | :||          ||                  |
Db   2533 HELTTSSENVPKESESTTTSSESSKPSQEPAGILTSTVVVPTSSVSLITASEIEAITSNT 2592

Qy    716 P--KGTAP--TTPK-------EPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKEL 764
          |  :|  |  |:||         |:  |  ||: :| : |  ||:    ||     : :
Db   2593 PFKQGRTPITTSPKSLVKSTTSPSTVTSSEPSESTKRTTVSTTVSTTTPTEETTTSESLI 2652

Qy    765 APTTTKGPT-STTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPT 823
            || |||    ||||| :|: | |   : |: |        |    : | || :
Db   2653 LTAAPSKPTESTTESSEAPTTPAKTSETKPSNVSSTSRKSTENVETSTSQSGSLESSTMS 2712

Qy    824 TTKEPTTIHKSPDESTP--ELSAEPTPKALE--NSPKEPGVPTTKTPAATK-----PEMT 874
          :|         |: : |   :|:| :   ||  :|   |     | : |: |      ||
Db   2713 STS------SEPETNAPAVTVSSEASSTTLEENSSTSSP----TSSEASVKLSSLFPESI 2762

Qy    875 TTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTT-------------- 920
          |:     : |     |   | : :  :| :| : : :| :|  :||
Db   2763 TSEAVTVSSR----APAEITMSSESHREISTVSSEPSEPEIPLSTTVSPNVVTASSIPSE 2818

Qy    921 -QVTSTTTQDTTP-FKITT-----LKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRA 973
           : |: |  :||  :: |     |  :   |     :: || : :   |  :  |
Db   2819 EPILSSVTSSSTPRVRLITGTPDDLIVSVTVPSHGNRRQNITASSV---PSNSTSPIILP 2875

Qy    974 TNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPK---TTPTPRKMTSTMPELN----- 1025
          : |  |||:|   |    || :|   :  | :: |    ||| |       |  |
Db   2876 SES-LTTPQPPPTTTTTAKPATTSGKRGPPSIQPPAEMFTTPAP------PPPSNGGYGE 2928

Qy   1026 PTSRIAEAMLQTTT 1039
           |::  | :  |||
Db   2929 ETNQEEEQVTSTTT 2942


RESULT  14
I51618
nucleolar phosphoprotein - African clawed frog
C;Species: Xenopus laevis (African clawed frog)
C;Date: 13-Sep-1996 #sequence_revision 13-Sep-1996 #text_change 24-Sep-1999
C;Accession: I51618; S57757
R;Cairns, C.; McStay, B.
J. Cell Sci. 108, 3339-3347, 1995
A;Title: Identification and cDNA cloning of a Xenopus nucleolar phosphoprotein, xNopp180
A;Reference number: I51618; MUID:96019267
A;Accession: I51618
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: mRNA
A;Residues: 1-990 <CAI>
A;Cross-references: EMBL:X88927; NID:g895920; PIDN:CAA61368.1; PID:g895921
C;Genetics:
A;Gene: xNopp180
C;Superfamily: nucleolus-cytoplasm shuttle phosphoprotein
C;Keywords: phosphoprotein


  Query Match             8.8%;  Score 643.5;  DB 2;  Length 990;
  Best Local Similarity   24.2%;  Pred. No. 2e-20;
  Matches  246;  Conservative  148;  Mismatches  424;  Indels  199;  Gaps   40;

Qy      74 SPPSSKKAPPPSG--------ASQTIKSTTKRSPKPPNKKKT------KKVIESEEITEE 119
           || : |: || :|         | : : ::    :|| ||:       | |::: :  :
Db      57 SPDAKKRKPPANGLPKKKSAKESSSSEDSSSEEDEPPAKKRAQPAGGKKPVVKAVQPKKA 116

Qy     120 HSVSENQESSSSSSSSSSSSTIWK-IKSSKNSAANRELQKKLKVKDNKKNRT------KK 172
           | ||:    | |   :       :  :: | :|     ||| |   :: : :       ||
Db     117 KSSSEDSSDESDSEEETKKPPAKRPAQTPKVAAVKTPTQKKAKSSSSESSSSEDEASKKK 176

Qy     173 KPTPKPP---VVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPRPSLPPN 229
             :|  | |    | :||   :||     | |:|::: :    : |    |         ||
Db     177 QPVIKVPPKQAVVKAGLASNNG----KTADSSSSEDSDSPPAKKTAATK-------TPP- 224

Qy     230 SDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTS-- 287
             |   |:     : :|   | :: ':   |:| ::||  :|    :           |||
Db     225 --TKPATAAKPQAKKTAGKKSSSREDSSDSSDEEQKTAKSKPKPDVYSAVP---PPTSVS 279

Qy     288 --KVLAKPTPKAETTTKGPALTTPKEPTPTTPKE--PASTTPKEPTPTTIKSAPTTPKEP 343
             | |::|  ||:  :   : :: :|  |    :  ||      | |   |:  :|  |
Db     280 KKKTLSQPGTKAKPESSDSSDSSDEEEQPAKKAKIVPAKAAASAPKPLAKKAETSTDSES 339

Qy     344 APTT-----------TKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKE-PAPTTTKSAP 391
             ::             |:||   | ||       ||      |     |  ||   : |:
Db     340 DSSSEDEKKSSVKLGVKAAP--KKAPA-------APDAKSTPVAAAKKSAPAKKASSSSD 390

Qy     392 TTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPA-PTTPKEPAPTAPK 450
           :          || |  | ||| : |  ||   |||  |   ||:| || | |||
Db     391 SDSSSNEETTTKPAAKTTPAKSA-ATPTSKTPTNGK-ATPTSKTPAKPGTPKTSTAKKDS 448

Qy     451 KPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTT 510
             : ::        |  :||  ||    |  ||     || :|: ||: | ||   |
Db     449 SSSDSSDSSSDEETTTKPAAKTTPAKSAATPTSKTPTNSKATPTSKKTPAKPGTPKTSAA 508

Qy     511 PKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEP-APTTPKEPAPTTTKKPAPTAPKE 569
             |: | : : : :    | ||    |: | ||| :| | ||| :||       | | | :
Db     509 KKDSSSSDSSDSSSDEKKTPA----KRAAKTTPAKPAAKTTPAKPA-------AKTTPAK 557

Qy     570 PAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 629
           ||                          | :|| |  ||  |    ::  : : :: |:   :
Db     558 PA----------------------AKSTPGKQVPTKKE----SSSSDSSDSSSEDEKKS 590

Qy     630 TPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKK 689
           : | |   |   |  | |   |    :| |   | :: :    :  | |   | |
Db     591 SAKPAVKTT---PGKATSK---PVVASKPVP-AKKASSSSDSDSSEEETTKTTKPLTKLS 643

Qy     690 PAPKELAPTTTKEPTSTTSD-------KPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKG 742
           || | | |   :  : ::||        |||     |   |   | ||     : : :
Db     644 PAVKTLPPKKAESSSDSSSDSDSEKKTKPAKPPAKSATPVNTKAPAQNKASKASCSDSDS 703

Qy     743 TAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPK 802
           ::    |  | | | || |  ||   | | :   |||: ::  :::    |:     ||
Db     704 SSEEEGKSKQP-TGKSPAAKATAP-PKKNPVAVNKDKPSSSSSSDSSGDDEKQ----KPK 757

Qy     803 KPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPT 862
           : |                                : |      :|:|||   :|  |    :
Db     758 QAA--------------------------AAKDVKQGAKAAKPTPKKAASSSSE---DS 787

Qy     863 TKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQV 922
           :    :| : | ||  |:      |||:   || | :   :: :|   :     :||::
Db     788 SSDEDVSKAKKTNTAVSKSPV----TTPKAVPAAKKESSSESSDSEDEKQGGKNTSTTKI 843

Qy     923 TSTTTQDTTPFKITTLKTTTLAPKVTTTK--KTITTTEIMNKPEETA-KPKDRATNS--- 976
             ::|                  ||    :  :  :::|   |   |: | |  :| :
Db     844 ANST-----------------PKAAAAECSEESSSSEDEGKANGTSGKRKRESTGNAEC 885

Qy     977 KATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTP-RKMTSTMPELNPTSRIAE 1032
           :| ||:         |    | | | |:| | :   || |::     |:||  |:|:
Db     886 EAVTPE--------NKKLKAKSPNTFPKVNKKELKNTPFRRVVEEDIEINP--RMAD 932


RESULT  15
T11622
extensin class 1 precursor - cowpea
C;Species: Vigna unguiculata (cowpea)
C;Date: 16-Jul-1999 #sequence_revision 16-Jul-1999 #text_change 17-Nov-2000
C;Accession: T11622; S54155
R;Arsenijevic-Maksimovic, I.; Broughton, W.J.; Krause, A.
Mol. Plant Microbe Interact. 10, 95-101, 1997
A;Title: Rhizobia modulate root-hair-specific expression of extensin genes.
A;Reference number: Z17301; MUID:97155574
A;Accession: T11622
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
```

```
Qy    478 ------------------------------PTTPKEPAPTTTKSAPTTT--KEPAPTTT 504
Db   2286 GLVCRNREQVGKFKMCFNYEIRVFCCNYGHCPSTPATSSTATPSSTPGTTWILTKLTTTA 2345

Qy    505 KSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPA- 563
Db   2346 TTTESTGSTATPSSTQGPPAGTPH--VSTTATTPTVTSSK----ATPFSSPGTATALPAL 2399

Qy    564 ------PTAPKEPA---------------PTTPKETAPTTPKKLTP-------------- 588
Db   2400 RSTATTPTATSFTAIPSSSLGTTWTRLSQTTTPMATMSTATPSSTPETVHTSTVLTTTAT 2459

Qy    589 --------TTPEKLAPTTPEKPAPTT-------PEELAPTTPEEP-----TPTTPEEPAP 628
Db   2460 TTGATGSVATPSSTPGTAHTTKVPTTTTTGFTVTPSSSPGTARTPPVWISTTTTPTTSGS 2519

Qy    629 TTPKAAAPNTPKEPAP-TTPKEPAPT----TPKEPAPT--TPKE--TAPTTPKGTAPTTL 679
Db   2520 TVTPSSIPGTTHTPTVLTTTTQPVATGSMATPSSSTQTSGTPPSLITTATTITATGSTT- 2578

Qy    680 KEPAPTTPKKPAPKELAPT-TTKEPTS----------TTSDKPAPTTPKGT-----APTT 723
Db   2579 -NPSSTPGTTPIPPELTTTATTPAATSSTVTPSSALGTTHTPPVPNTTATTHGRSLSPSS 2637

Qy    724 P-------------------KEPAPTTPKEPAPTTPKGTAPT-TLKEPAP---TTPKKP 759
Db   2638 PHTVRTAWTSATSGTLGTTHITEPSTGTSHTPAATTGTTTTSTPALSSPHPSSRTTESPP 2697

Qy    760 APKELAPTTTKGPTSTTSDKPAPTTPKETAPTT--PKEP---------------------- 796
Db   2698 SP----GTTTPGHTTATSRTTATATPSKTRTSTLLPSQPTSAPITTVVTTGCEPQCAWSE 2753

Qy    797 ------------------------------------------------------------ 796
Db   2754 WLDYSYPMPGPSGGDFDTYSNIRAAGGAVCEQPLGLECRATAQPGVPLGELGQVVECSLD 2813

Qy    797 ------------------------------APTTPKKPAPTTPETPPPTT---SEVSTP 822
Db   2814 FGLVCRNREQVGKFKMCFNYEIRVFCCNYGHCPSTPATSSTAMPSSTPGTTWILTELTTT 2873

Qy    823 TTTKEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTK----------------TP 866
Db   2874 ATTTASTGSTATP-SSTP--GTAPPPKVLTSPATTPTATSSKATSSSSPRTATTLPVLTS 2930

Qy    867 AATK--------------------PEMTTTAKDKTTERDLRTTPETT-TAAPKMTKETAT 905
Db   2931 TATKSTATSVTPIPSSTLGTTGTLPEQTTTPVATMSTIHPSSTPETTHTSTVLTTKATTR 2990

Qy    906 TTEKT-------------TESKITATTTQVTS-TTTQDTTP------FKITTLKTTTLAP 945
Db   2991 ATSSTSTPSSTPGTTWILTELTTAATTTAGTGPTATPSSTPGTTWILTELTTTATTTAST 3050

Qy    946 KVTTT-KKTITTTEIMNKPEE----TAKPKDRAT--NSKATTPKPQ------KPTKAPKK 992
Db   3051 GSTATLSSTPGTTWILTEPSTTATVTAPPGSTATASSTQATAGTPHVSTTATTPTVTSSK 3110

Qy    993 PTSTKKPKT---MPRVRKPKTTPTPRKMT----------------------------ST 1020
Db   3111 ATPSSSPGTATALPALRSTATTPTATSFTAIPSSSLGTTWTRLSQTTTPTATMSTATPSS 3170

Qy   1021 MPELNPTSRIAEAMLQTT------TRPNQTPNSKLVEVNPKSEDAG--GAEGETPHMLLR 1072
Db   3171 TPETVHTSTVLTTTATTTGATGSVATPSSTPGTAHTTKVPTTTTTGFTATPSSSPGTALT 3230

Qy   1073 PHVFMPEVTPDMDYLPRVPNQGIIINPMLSDETNICNGKPVDGLTTLRNGTLVAFRGHYF 1132
Db   3231 PPVWISTTTTPTTTTP--TTSGSTVTPSSIPGTTHTARVLTTTTTTVATGSM-------- 3280

Qy   1133 WMLSPFSPPSPARRITEVWGIPSPIDTVFTRCNCEGKT 1170
Db   3281 ------ATPSSS---TQTSGTPPSLTTTATTITATGST 3309
```

```
RESULT  13
T34513
hypothetical protein ZK783.1 - Caenorhabditis elegans
C;Species: Caenorhabditis elegans
C;Date: 29-Oct-1999 #sequence_revision 29-Oct-1999 #text_change 29-Oct-1999
C;Accession: T34513
R;Favello, A.; Vaudin, M.
submitted to the EMBL Data Library, August 1994
A;Description: The sequence of C. elegans cosmid ZK783.
A;Reference number: Z21536
A;Accession: T34513
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-3507 <FAV>
A;Cross-references: EMBL:U13646; PIDN:AAC24418.1; GSPDB:GN00021; CESP:ZK783.1
A;Experimental source: strain Bristol N2; clone ZK783
C;Genetics:
A;Gene: CESP:ZK783.1
A;Map position: 3
A;Introns: 14/1; 48/2; 84/1; 196/3; 303/1; 381/1; 586/1; 605/1; 1175/3; 1207/1; 1409/
 3504/1


  Query Match             8.9%;  Score 647.5;  DB 2;  Length 3507;
  Best Local Similarity   23.3%;  Pred. No. 4.9e-20;
  Matches  283;  Conservative  182;  Mismatches  496;  Indels  253;  Gaps   52;

Qy     27 LSSCAGRCGEGYSRD-ATC-----------NCDYNCQHYMECCPDFKRVCTAAEVHNPT 73
Db   1781 MGSCGCKCMAGYTGDGATCIKIEEEPKSDKTAC---TDEWSRLCELEKKQCTVDEEEVPQ 1837

Qy     74 SPPSSKKAPPPSGASQTIKSTTKRSPKPPNKKKTKKVIESEEI-TEEH----------- 120
Db   1838 CGACLPGHHPINGTCQSLQI----SGLCAQKNDCNKHAECIDIHPDSHFCSCPDGFIGDG 1893

Qy    121 SVSENQESSSSSSSSSSSSTIWKIKSSKNSAANRELQKKLKVKDNKKNRTKKKPTPKPPV 180
Db   1894 MICDDVDECNNAGMCDDENT--KCENTIGSFNCVCLEGFKKVDEKCVVDEKKQPNREKIE 1951

Qy    181 VDEAGSGLDN-GDFKVTT----PDTSTTQHNKVSTSPKITTA-------KPINPRPSLPP 228
Db   1952 IDEENSSSSNSGQEKPTTKGIVSSTSATSSESTTAEPHVTTSISSTTSTKDMTSSKS-PE 2010

Qy    229 N---SDTSKETSLTVNKETTV-ETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLA 284
Db   2011 NVTMSSESPEVSTSSSKSTTASETTVSSTPSESSSSEAPLTSSPATTTEVITESSVKSTT 2070

Qy    285 P----TSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKS-APTT 339
Db   2071 PKEESSSEITVKLSSKSPEVTESSVKSSPSTPS-TTSQSVTSTVPETSKSTVLSSEAPVT 2129

Qy    340 PKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAP 399
Db   2130 STSPTEVHT-SSETKPSLSASSTTGDTNSTTPSTSSLASVKSTSAPEGTSASVAPVKLSS 2188

Qy    400 TTPKKPAPTTPK-EPAPTTPKEPTPTTPKEPAPTTKEP-APTTPKEPAPTAPKKPAPTTP 457
Db   2189 LSPDVSQPSTKTFDATESSTVQASETSSGTSVKSTSEPESHVTKLSITSSNPSSSVPVTS 2248

Qy    458 KEPAPTTPKEPAPTTTKEPSPTTP--KEPAPTTTKSAPTTTKEPAPTTTKSAP------T 509
Db   2249 PKSTPTVPE-----STEQPTSTTPSGQSLTPMNSNSEVLTTSEPHVLSSSLSPDVSQSST 2303

Qy    510 TPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTT----------- 558
Db   2304 TPNNLSESSTVE----TPKTSSEVSLNSEEPSTTEAPTTLSPDILSTTTNNLSQSSTVST 2359

Qy    559 ----------TKKPAPTAPK----------EPAPTTPKETAPTTPKKLTPTTPEKLAPTT 598
Db   2360 EDRSEISSENSEKPT-SAPELVTSSVTHVASSSPDVPTES--SEPDDLTGSSTENIPEAS 2416

Qy    599 PEKPAPTTPEELAPTTPEEPTPTTPEEP-----------APTTPKAAAPNTPKEPAPTT- 646
```

```
  Query Match             9.1%;  Score 661.5;  DB 2;  Length 7962;
  Best Local Similarity   24.3%;  Pred. No. 2.9e-20;
  Matches  266;  Conservative  121;  Mismatches  480;  Indels  229;  Gaps   50;

Qy      58 PDFKRVCTAAEVHNPTSPPSSKKAPPPSGASQTIKSTTKRSPKPPNKKKTKKVI--ESEE 115
Db    6853 PEIKKKVTEKKV----VIPKKEEAPP-----------AKVPEVPKKVEEKRIILPKEEE 6896

Qy     116 ITEEHSVSENQESSSSSSSSSSSTIWKIKSSKNSAANRELQKKLKVKDNKKNRTKKKPT 175
Db    6897 VLPVEVTEEPEEEPISEE---------EIPEEPPSIEEVEEVAPPRVPEVIKKAVPEAPT 6947

Qy     176 PKPPVVDEAGSGLDN--GDFKVTTP-DTSTTQHNKVSTSPKITTAKP--INPRPSLPP-- 228
Db    6948 PVPKKVEAPPAKVSKKIPEEKVPVPVQKKEAPPAKVPEVPKKVPEKKVLVPKKEAVPPAK 7007

Qy     229 -NSDTSKETSLTVNKETTVETK---ETTTTNKQTSTDGKE------------------- 264
Db    7008 GRTVLEEKVSVAFRQEVVVKERLELEVVEAEVEEIPEEEEFHEVEEYFEEGEFHEVEEFI 7067

Qy     265 --KTTSAKETQSIEKT-------SAKDLAPTSKVLAKPTPKAETTTKGPALT-----TPK 310
Db    7068 KLEQHRVEEEHRVEKVHRVIEVFEAEEVEVFEKPKAPP--------KGPEISEKIIPPKK 7119

Qy     311 EPTPTTP-KEPASTTPKEPTPTTIKSAPTTPKEP-APTTTKSAPTTPKEPAPTTTKEPAP 368
Db    7120 PPTKVVPRKEPPAKVPEVPKKIVVEEKVRVPEEPRVPPTKVPEVLPPKEVVP-------- 7171

Qy     369 TTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKK----PAPTTPKEPAPTTPKEPTPT 424
Db    7172 -EKKVPVPPAKKPEAP-----PPKVPEAPKEVVPEKKVPVPPPKKPEVPPTKVPEVPKAA 7225

Qy     425 TPKEPAPTTKEPAPTTP-------KEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPS 477
Db    7226 VPEKKVPEAIPPKPESPPPEVFEEPEESPSAPPK-KPEVPPVRVPEVPKEVVPEKKVPAA 7284

Qy     478 PTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKK 537
Db    7285 P--PKKPEVTPVK-VPEAPKEVVPEKKVPVP-PPKKPEVPPTK-----VPEVPKVAVPEK 7335

Qy     538 PAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLT----PTTPEK 593
Db    7336 KVP----EAIPPKPESPPPEVFEEPEEVALEEPPAEVVEEPEPAAPPQVTVPPKNPVPEK 7391

Qy     594 LAPTT----PEKPAPTTPEELAPTTPEEPTP-TTPEEPAPTTPKAAAPNTPKEPAP---- 644
Db    7392 KAPAVVAKKPELPPVKVPEVPKEVVPEKKVPLVVPKK--PEAPPAKVPEVPKEVVPEKKV 7449

Qy     645 TTPKEP------APTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPT 698
Db    7450 AVPKKPEVPPAKVPEVPKKPV----LEEKPAVP---VPERAESPPPEVYEE--PEEIAP- 7499

Qy     699 TTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKK 758
Db    7500 --EEEIAPEEEKPVPVAEE----EEPEVPPPAVPEEPKKIIPEKKVP-VIKKPEAPPPKE 7552

Qy     759 PAPKELAPTTTKGPTSTTSDKPAPTTPKETA-------PTTPKEPAPTTPKKP--APTTP 809
Db    7553 PEPEKV---IEKPKLKPRPPPPPPAPPKEDVKEKIFQLKAIPKKKVPENPQVPEKVELTP 7609

Qy     810 ETPPPTTSEVS--TPTTTKEP-------TTIHKSPDESTPELSAEPTPKALENSPKEPGV 860
Db    7610 LKVPGGEKKVRKLLPERKPEPKEEVVLKSVLRKRPEEEEPKVE----PKKLE-KVKKPAV 7664

Qy     861 PTTKTPAATKP----EMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKIT 916
Db    7665 P---EPPPPKPVEEVEVPTVTKRERKIPEPTKVPEIKPAIPLPAPEPKPKPEA------- 7714

Qy     917 ATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNK------PEETAKPK 970
Db    7715 ----------------EVKTIKPPPVEPEPTPIAAPVTVPVVGKKAEAKAPKEEAAKPK 7757

Qy     971 DRATNSKATTPKPQKPTKAPKKPTS--TKKPKTMPRVRKPKTTPTP-----RKMTSTMPE 1023
Db    7758 GPIKGVPKKTPSPIEAERRKLRPGSGGEKPPDEAPFTYQLKAVPLKFVKEIKDIILTESE 7817

Qy    1024 LNPTSRIAEAMLQTTT 1039
Db    7818 FVGSSAIFECLVSPST 7833


RESULT  12
T45025
mucin MUC5B, tracheobronchial [imported] - human (fragment)
C;Species: Homo sapiens (man)
C;Date: 21-Jan-2000 #sequence_revision 21-Jan-2000 #text_change 21-Jul-2000
C;Accession: T45025
R;Desseyn, J.L.; Guyonnet-Duperat, V.; Porchet, N.; Aubert, J.P.; Laine, A.
J. Biol. Chem. 272, 3168-3178, 1997
A;Title: Human mucin gene MUC5B, the 10.7 kb large central exon encodes various alter
A;Reference number: Z22899; MUID:97166151
A;Accession: T45025
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-3570 <DES>
A;Cross-references: EMBL:Z72496; NID:g1834502; PIDN:CAA96577.1; PID:g1834503
A;Experimental source: placenta
C;Genetics:
A;Gene: MUC5B


  Query Match             8.9%;  Score 649;  DB 2;  Length 3570;
  Best Local Similarity   22.5%;  Pred. No. 4.3e-20;
  Matches  373;  Conservative  111;  Mismatches  584;  Indels  590;  Gaps   64;

Qy      62 RVCTAAEVHNPTSPPSSKKA--PPPSGASQTIKSTTKRSPKPPNKKKTKKVIESEEITEE 119
Db    1693 RTATTLPVLTSTATKSTATSFTPIPSSTLGTTGTSQNRPPHPMATMSTIHPSSTPETTHT 1752

Qy     120 HSV-----SENQESSSSSSSSSSSTIWKIKSSKNSAANRELQKKLKVKDNKKNRT--KK 172
Db    1753 STVLTTKATTTRATSSMSTPSSTPGTTW-ILTELTTAATTTAALPHGTPSSTPGTTWILT 1811

Qy     173 KPTPKPPVVDEAGS-------GLDNGDFKV-----TTPDTSTTQHNKVSTSPKITTAKPI 220
Db    1812 EPSTTATVTVPTGSTATASSTRATAGTLKVLTSTATTP-TVISSRATPSSSPGTATA--- 1867

Qy     221 NPRPSLPPNSDTSKETSLTVNKETTVET-----KETTTTNKQTSTDGKEKTTSAKETQSI 275
Db    1868 --LPALRSTATTPTATSVTAIPSSSLGTAWTRLSQTTTPTATMSTATPSSTPETVHTSTV 1925

Qy     276 EKTSAKDLAPTSKVLAKPTPKAETTTK----------------GPALTTP----KEPTP 314
Db    1926 LTTTATTTRTGSVATPSSTPGTAHTTKVPTTTTTGFTATPSSSPGTALTPPVWISTTTTP 1985

Qy     315 T------TPKE-----------------------PASTTPKEPTPTTIKSAPTTPKEP 343
Db    1986 TTRGSTVTPSSIPGTTHTATVLTTTTTTVATGSMATPSSSTQTSGTPPSLTTTATTITAT 2045

Qy     344 APTTTKSAP--TTPKEPAPTTT-KEPAPT----TPKEPAPTTTKEPAPTTTKSA------ 390
Db    2046 GSTTNPSSTPGTTPIPPVLTTTATTPAATSSTVTPSSALGTTHTPPVPNTTATTHGRSLP 2105

Qy     391 PTTP--------------------KEPAPTTPKKPAPTT----PKEPAPTTPKEPTPTTP 426
Db    2106 PSSPHTVPTAWTSATSGILGTTHITEPSTGTSHTPAATTGTTQPSTPALSSPHPSSRTTE 2165

Qy     427 KEPAPTTKEPAPT--TPKEPAPTAPKKPAPTTPKEPAPTT-PKEPAPTTTKEPS------ 477
Db    2166 SPPSPGTTTPGHTRGTSRTTATATPSKTRTSTLLPSSPTSAPITTVVTTGCEPQCAWSEW 2225

Qy     478 ---------------------------------------------------------- 477
Db    2226 LDYSYPMPGPSGGDFDTYSNIRAAGGAVCEQPLGLECRAQAQPGVPLRELGQVVECSLDF 2285
```

```
Biochem. Biophys. Res. Commun. 165, 644-649, 1989
A;Title: Sequence analysis of the 5' region of the human DF3 breast carcinoma-associated
A;Reference number: A36735; MUID:90088473
A;Accession: A36735
A;Molecule type: mRNA
A;Residues: 1-142,'Q',144-162,'Q',164-168 <ABE>
A;Cross-references: EMBL:M31823; NID:g181542; PIDN:AAA35757.1; PID:g181543
R;Masuzawa, Y.; Miyauchi, T.; Hamanoue, M.; Ando, S.; Yoshida, J.; Takao, S.; Shimazu, H
J. Biochem. 112, 609-615, 1992
A;Title: A novel core protein as well as polymorphic epithelial mucin carry peanut agglu
A;Reference number: JX0235; MUID:93123189
A;Accession: PX0066
A;Molecule type: mRNA
A;Residues: 998-1011,'ES',1014-1017;1018-1032,'T',1034-1037;1038-1057 <MAS>
A;Experimental source: gastric carcinoma cell
R;Zrihan-Licht, S.; Baruch, A.; Elroy-Stein, O.; Keydar, I.; Wreschner, D.H.
FEBS Lett. 356, 130-136, 1994
A;Title: Tyrosine phosphorylation of the MUC1 breast cancer membrane proteins cytokine r
A;Reference number: S51026; MUID:95080414
A;Contents: annotation
A;Note: undetermined tyrosine residues in the carboxyl-terminal non-repetitive region ar
C;Comment: This protein is length polymorphic. Individuals may have between 21 and 125 c
 partial repeats. The repeat shown is defined by SmaI nuclease sites.
C;Comment: Serine and threonine residues in the tandem repeat domain are extensively gly
C;Comment: For an alternative splice form without a tandem repeat domain, see PIR:S48146
C;Genetics:
A;Gene: GDB:MUC1; PUM
A;Cross-references: GDB:120705; OMIM:158340
A;Map position: 1q21-1q23
A;Introns: 20/1; 62/3; 1165/3; 1184/2; 1230/1; 1270/3; 1320/3
C;Superfamily: polymorphic epithelial mucin
C;Keywords: alternative splicing; duplication; glycoprotein; phosphoprotein; polymorphis
F;1-1344/Product: mucin 1 precursor, splice form A #status predicted <PREA>
F;1-62/Region: mucin 1 amino-terminal non-repetitive
F;1-23/Domain: signal sequence #link PREA #status predicted <SIGA>
F;1-19,29-32/Domain: signal sequence #link PREB #status predicted <SIGB>
F;1-19,29-1344/Product: mucin 1 precursor, splice form B #status predicted <PREB>
F;1-19,29-212,1033-1344/Product: mucin 1 precursor, epithelial tumor antigen splice form
F;138-1017/Region: 20-residue repeats (GSTAPPAHGVTSAPDTRPAP)
F;1143-1344/Region: mucin 1 carboxyl-terminal non-repetitive
F;1245-1272/Domain: transmembrane #status predicted <TRM>
F;1046,1064,1118,1144,1222/Binding site: carbohydrate (Asn) (covalent) #status predicted
F;1213/Binding site: phosphate (Tyr) (covalent) #status predicted

  Query Match             9.2%;  Score 672;  DB 1;  Length 1344;
  Best Local Similarity   27.8%;  Pred. No. 1.7e-21;
  Matches  310;  Conservative   78;  Mismatches  472;  Indels  256;  Gaps   56;

Qy    238 LTVNKETTVETKETTTTNK--QTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 294
Db     15 LTVLTATTAPKPATVVTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 74

Qy    295 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTP----------KE 327
Db     75 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVTSAPDNK 134

Qy    328 PTPTT-------IKSAPTTPKEPAPTTT-------KSAPTTPKEPAPTTTKEP------A 367
Db    135 PAPGSTAPPAHGVTSAPDT--RPAPGSTAPPAHGVTSAPDT--RPAPGSTAPPAHGVTSA 190

Qy    368 PTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTP----------KKPAPTTPKEP---- 413
Db    191 PDT--RPAPGSTAPPAHGVT-SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGV 247

Qy    414 --APTTPKEPTPTTP----------KEPAPTTKEP-------APTTPKEPAPTAPK---- 450
Db    248 TSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGV 307

Qy    451 KPAPTTPKEPAPTTPK----EPAPTTTKEPSPTTPKEPAPTT---TKSAPTTTKEPAPTT 503
Db    308 TSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGV 367

Qy    504 TKSAPTTPKEPSPTTTKEP------APTTPKEPAPTTPK----KPAPTTPKEPAPTTPK- 552
Db    368 T-SAPDT--RPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPA 424

Qy    553 ---EPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPE- 608
Db    425 HGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPA 484

Qy    609 ---ELAPTTPEEPTPTTP----------EEPAP--TTPKA----AAPNTPKEPAPTTPK- 648
Db    485 HGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPA 544

Qy    649 ---EPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTP----------KKPAPKEL 695
Db    545 HGVTSAPDTRPAPGSTAP----PAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGST 600

Qy    696 APTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTPKEPAPTTPKG----TAPTTLK 749
Db    601 AP-PAHGVTSAPDTRPAPGSTAPPAHGVTS----APDTRPAPGSTAPPAHGVTSAPDTRP 655

Qy    750 EPAPTTP----------KKPAPKELAPTTTKGPTSTTSDKPAP--TTPK----ETAPTTP 793
Db    656 APGSTAPPAHGVTSAPDTRPAPGSTAP-PAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTR 714

Qy    794 KEPAPTTP----------KKPAPTTPETPPPTTSEVSTPTTTKEPTT-------IHKSPD 836
Db    715 PAPGSTAPPAHGVTSAPDTRPAPGS--TAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPD 772

Qy    837 ESTPELSAEPTPKALENSPKEPGVPTTKTPAA--------TKPEMTTTAKDK---TTERD 885
Db    773 TRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPD 832

Qy    886 LRTTPETT-------TAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTP--FKIT 936
Db    833 TRPAPGSTAPPAHGVTSAP----DTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT 888

Qy    937 TLKTTTLAPKVTTTKKTITTT--EIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPT 994
Db    889 SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAP--GSTAPPAHG 946

Qy    995 STKKPKTMPRVRKPKTTPTPRKMTSTMPELNP----TSRIAEAMLQT-TTRP---NQTPN 1046
Db    947 VTSAPDTRP---APGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPP 1003

Qy   1047 SKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSDETN 1106
Db   1004 AHGVTSAPDTRPAPGSTAPPAH----------GVTSAPDNRPALGSTA----PPVHNVTS 1049

Qy   1107 ICNGKPVDGLTTLRNGTLVAFRGHYFWMLSPFSPPS 1142
Db   1050 ASGSASGSASTLVHNGTSARATTTPASKSTPFSIPS 1085


RESULT  11
I38346
elastic titin - human (fragment)
C;Species: Homo sapiens (man)
C;Date: 29-May-1998 #sequence_revision 29-May-1998 #text_change 21-Jul-2000
C;Accession: I38346
R;Labeit, S.; Kolmerer, B.
Science 270, 293-296, 1995
A;Title: Titins: giant proteins in charge of muscle ultrastructure and elasticity.
A;Reference number: A57430; MUID:96026330
A;Accession: I38346
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: mRNA
A;Residues: 1-7962 <RES>
A;Cross-references: EMBL:X90569; NID:g1017426; PIDN:CAA62189.1; PID:g1017427
C;Genetics:
A;Gene: GDB:TTN
A;Cross-references: GDB:127867; OMIM:188840
A;Map position: 2q31-2q31
```

```
Qy      2 AWKTLPIYLLLLLSVFVIQQVSSQDLSSCAGRCGEGYSRD----------ATCNCDYNC 50
Db      4 AW-VVSFAFLILGNVVQSSSLLSKTINS------DG-SRDFKIVKHIKKNTCTCSC---- 51

Qy     51 QHYMECCPD-----FKRVCTAAEVHNPTSP--PSSKKAPPPSGASQTIKSTTKRSPKPPN 103
Db     52 ----KCVPDAPSNPFDVSTTISSINNDNVDIGPSGDSNPTGSSWFQEIEAT--------- 98

Qy    104 KKKTKKVIESEEITEEHSVSENQESSSSSSSSSSSTIWKIKSSKNSAANRELQKKLKVK 163
Db     99 -------VGGQTVKSEHNIDSSVEVEKKVTTSTDAST---------------------- 128

Qy    164 DNKKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPR 223
Db    129 TNAPTTGKDSTTPE----------IITGIVVINSKSESVTD---MSTTRFSTTLSP---- 171

Qy    224 PSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDL 283
Db    172 ----------TTELLTSPETLVSTDSSTST-EQTSPDNTTEIASPMETNTTTEATTTSV 219

Qy    284 APTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEP 343
Db    220 EPSVSTLAS---EDETTVTAIAEST------TTVIAEVSTTTEEPTTT------------ 258

Qy    344 APTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTP- 402
Db    259 AESTTKKSTT----KAPATTEEPTPTTTEE--VTTTEAETSTTTSSETSTEK---PTTPL 309

Qy    403 ---KKPAPTTPKEPAPTTPKEP-TPTTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPK 458
Db    310 IDNKIAGPATGK---PETTHFPVTGTTPN--FDTATETPFVAKSEDKMTLSKTAATETTQ 364

Qy    459 EPAPTTPKEPAPTTTKE-----PSPTTPKEPAPTTTKSAP-------TTTKEPAPT---- 502
Db    365 QTTEVT-DGPEKETTKNVSIEIPITTVPLVETTSTTSTASKESDGFHTTLKLKVTTADSD 423

Qy    503 TTKSAPTTPKEPSPTTTKE---PAPT-------TPK------EPAPTTPKKPAP------ 540
Db    424 STESATTVKPFNEETTTKSHVVPKPTKKGTVKVTPKLELSFDEPTEIT-KAPHPGKLLEK 482

Qy    541 --------------TTPKE--------------PAPTTPKEPAP----TTTKKPAPTA- 566
Db    483 KTYHFVLSDNFARYSEAKENDDYNHLDYNHYREAKEPTTTEEESSTTEEVTTTEEPANTGN 542

Qy    567 -PKEPAPTT---PKETAPTTPKKLTPTTPEKLA---PTTPEKPAPTTPEELAPTTPEEPT 619
Db    543 PPTTENPTTTEQPTSTAESTTTALPFTTEQTVTTEEPTTAEKSTATQ----KPTTTQESV 598

Qy    620 PTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTL 679
Db    599 ST---EKTSTTKKA---STTEE--PTTTDEPTTTT---ESSTTGKATTPELSTTSEETTT 647

Qy    680 KEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAP----TTPKGTAPTTPKEPAPTTPKEP 735
Db    648 TELKITTE--------GSTTTEEPTTTAIFAEASTGIITTDEETTSTTSTTPEITSTKE- 698

Qy    736 APTTPKGTAPTTLKEPAPTTP----------------------KKPAPKELAPTTTKGP 772
Db    699 -IVTESAITQTSVSVVESSTPRQLPERWKAIVNKFKHNLEVLKEKKRLLKEKESTSTTGS 757

Qy    773 ---------------TSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPE--TPPP 814
Db    758 DSSETTTVVAENIDEVTTTEKEKVVQTTPITTEKSTTQEETTTTTTTTTEKTTSKTTTEKP 817

Qy    815 TTSE-VSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEM 873
Db    818 TTSESATTETTTSEPST-----TEST--------------------TVDTSSATTEES 850

Qy    874 TTTAKDKTTERDLRTTPETTTA---------------------------------- 895
Db    851 STAAETTTTSAE---TSETTTSESAAFITGESPENTALQSSSQKSEENESSAEKPGARRD 907

Qy    896 -APKMTKETATTTEKTTESKITATTTQ-VTST---TTQDTTPFKITTLKTTTLAPKVT-- 948
```

```
Db    908 FVPKKHKTTVKPAETTSAVAASTTTTEPITTTEKSTTLETTPIEATTLNEVTGPAFVTGA 967

Qy    949 -TTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ----------------KPTKAP 990
Db    968 PVDETTINTLELLSK----------INNTQISQPKPTDISKTDALSSLISGLIGSFTKAP 1017

Qy    991 KKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELN-----------PTSRIAEA 1033
Db   1018 MAPTI------------HTTTDAAFVTATEASLNDGSDKKIIDEAQPTDEIRRA 1059
```

RESULT 10
A35175
mucin 1 precursor, repetitive splice form A [validated] - human
N;Alternate names: breast carcinoma-associated DF3 antigen; core protein KP39; episia
ncreatic mucin; polymorphic epithelial mucin (PEM)
N;Contains: mucin 1 precursor, epithelial tumor antigen splice form; mucin 1 precurso
C;Species: Homo sapiens (man)
C;Date: 20-Apr-2000 #sequence_revision 20-Apr-2000 #text_change 02-Jun-2000
C;Accession: A35175; B35175; A35886; A35887; S10572; S40293; A36735; PX0066; S10218;
R;Ligtenberg, M.J.L.; Vos, H.L.; Gennissen, A.M.C.; Hilkens, J.
J. Biol. Chem. 265, 5573-5578, 1990
A;Title: Episialin, a carcinoma-associated mucin, is generated by a polymorphic gene
A;Reference number: A35175; MUID:90202794
A;Accession: A35175
A;Molecule type: mRNA
A;Residues: 1-952,1033-1344 <LIG1>
A;Cross-references: GB:M32738; GB:J05288; NID:g182121; PIDN:AAA35804.1; PID:g182124;
A;Experimental source: splice form A
A;Note: GenBank entries HUMEPISIA1 and HUMEPISIA2 present only the amino-and carboxyl
A;Accession: B35175
A;Molecule type: mRNA
A;Residues: 1-19,29-952,1033-1344 <LIG2>
A;Cross-references: GB:M32739; GB:J05288; NID:g182126; PIDN:AAA35806.1; PID:g182129;
A;Experimental source: splice form B
A;Note: GenBank entries HUMEPISIB1 and HUMEPISIB2 present only the amino-and carboxyl
R;Gendler, S.J.; Lancaster, C.A.; Taylor-Papadimitriou, J.; Duhig, T.; Peat, N.; Burc
J. Biol. Chem. 265, 15286-15293, 1990
A;Title: Molecular cloning and expression of human tumor-associated polymorphic epith
A;Reference number: A35886; MUID:90368715
A;Accession: A35886
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 1-19,29-992,1033-1344 <GEN>
A;Cross-references: GB:J05581; NID:g188869; PIDN:AAA59876.1; PID:g188870
A;Note: GenBank entry HUMMUCAB includes one copy of the tandemly repeated sequence
R;Lan, M.S.; Batra, S.K.; Qi, W.N.; Metzgar, R.S.; Hollingsworth, M.A.
J. Biol. Chem. 265, 15294-15299, 1990
A;Title: Cloning and sequencing of a human pancreatic tumor mucin cDNA.
A;Reference number: A35887; MUID:90368716
A;Accession: A35887
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 1-19,29-1109,'S',1111-1339,'A',1341-1344 <LAN>
A;Cross-references: GB:J05582; NID:g189598; PIDN:AAA60019.1; PID:g189599
A;Note: GenBank entry HUMPANMU contains four fewer copies of the tandemly repeated se
R;Wreschner, D.H.; Hareuveni, M.; Tsarfaty, I.; Smorodinsky, N.; Horev, J.; Zaretsky,
Eur. J. Biochem. 189, 463-473, 1990
A;Title: Human epithelial tumor antigen cDNA sequences. Differential splicing may gen
A;Reference number: S10571; MUID:90276413
A;Accession: S10572
A;Molecule type: mRNA
A;Residues: 1-19,29-155,'P',157-175,'P',177-182,'A',184-212,1033-1344 <WRE>
A;Cross-references: EMBL:X52229; NID:g37053
R;Wreschner, D.H.
submitted to the EMBL Data Library, March 1990
A;Reference number: S40293
A;Accession: S40293
A;Molecule type: mRNA
A;Residues: 1-19,29-155,'P',157-175,'P',177-182,'A',184-212,1033-1037,'A',1039-1344 <
A;Cross-references: EMBL:X52229; NID:g37053; PIDN:CAA36478.1; PID:g37054
R;Abe, M.; Siddiqui, J.; Kufe, D.

```
Qy    725 -KEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPT 783
Db   1001 MSSPPPPEVKSPPPPAPVSSPPPPVKSPPPPAPVSSPP---PPVKSPPPPAPVSSPPPPV 1057

Qy    784 -TPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPEL 842
Db   1058 KSPPPPAPISSPPPPVKSPPPPAPVS--SPPPPVKSPPPPAPVSSPPPPIKSPPPPAPVS 1115

Qy    843 SAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPET 892
Db   1116 SPPPAP------VKPPSLP-PPAPVSSPPPVVTPAPPKKEEQSLPPPAES 1158


RESULT   8
T18535
high molecular mass nuclear antigen - chicken (fragment)
C;Species: Gallus gallus (chicken)
C;Date: 15-Oct-1999 #sequence_revision 15-Oct-1999 #text_change 15-Oct-1999
C;Accession: T18535
R;Shimada, K.; Harata, M.; Mizuno, S.
J. Cell Sci. 110, 3031-3041, 1997
A;Title: A nuclear matrix-associated high molecular mass nuclear antigen, HMNA, of chick
A;Reference number: Z18955; MUID:9803440
A;Accession: T18535
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: mRNA
A;Residues: 1-1151 <SHI>
A;Cross-references: EMBL:D88440; NID:d1177138; PID:d1025045; PIDN:BAA24137.1

  Query Match             9.5%;  Score 695;  DB 2;  Length 1151;
  Best Local Similarity   25.3%;  Pred. No. 1.6e-22;
  Matches  304;  Conservative  108;  Mismatches  466;  Indels  324;  Gaps   59;

Qy     67 AEVHNPTSPPSSKK------APPPSGASQTIKSTTKRSPKPPNKKKTKKVIESEEITEEH 120
Db     39 AELH-PAAPQPPPKWVPIGGAPPPPGTEPT-------PPSKPTDGADAAPKASAELTSPP 90

Qy    121 SVSENQESSSSSSSSSSSTIWKIKSSKNSAANRELQKKLKVKDNKKNRTKKKPTPKPPV 180
Db     91 PASPSPPDGPKAPSGAGEAEAGTPPPSQGPAG----------------------TPPP-- 126

Qy    181 VDEAGSGLDNGDFKVTTPDTSTTQHNKVSTS--PKITTAKPINPRPSLPPNSDTSKETSL 238
Db    127 -SQGAAGAPKGDGTAQPSGTKSGADGKPAAQDVPKATTA-ATEARP--------ASAASP 176

Qy    239 TVNKETTVETKETTTTNK--QTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPK 296
Db    177 TVPKATAEATAVTAASQSAPKAATDAAAVTAA---SQSAPKATV-EVKPAAAAVAKEAKA 232

Qy    297 AETTTKGPALTTPKEPTP-TTPKEPASTTPKEP----TPTTIKSAPTTPKEPAPTT---- 347
Db    233 VTAAAAAPKATAEAKPAPVTSPTIPCSSAEAKPLTAASPTASKA--TAEAKPVPATASLM 290

Qy    348 -------TKSAPTTPKEPAPTTTK----------------------EPAPTTPKEPAPT 377
Db    291 ATKVTAEAKPAPSPSVPKATTDTKAVTATAPKAGPDVKPAVAVCAEAKPAPPPPPQQLPK 350

Qy    378 TTKEPAPTTTKSAPTT------PKEPA-------PTTPKKPAPTTPKEPAPTTPKEPTPT 424
Db    351 AAAAAAPTGTELKPATAPPHGSPRANSHTVTVTPPNVPRAAAATVP--TAGAVPKASTGT 408

Qy    425 T----PKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEP---APTTTKEPS 477
Db    409 TPAAAPQQPVP---KAAPVTPPSPQQAVPRAATAAA----APVTPQQPVTKAATTTNATP 461

Qy    478 PTTPKEPAPTTTKSAPTTTKEPAP---TTTKSAPTTPKEPSPTTTKEP----APTTPKEP 530
Db    462 PPQPIPKAATTTTATPVTPQQPIPKAGTDAAPPPAVPKAPSDGRAATPGVPNAATDPPQKP 521

Qy    531 APTTPKKPAPTTPKEPAP---------TTPKEPAPT-TTKKPAPTAPK-------EPAPT 573
Db    522 PPTPQSVPSAVTEPKPQPRAAPPPSNEATPAVPSPSPNLKSPLPTIPKPVPLMALTPQPV 581

Qy    574 TPK---ETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELA-------PTTP-------- 615
Db    582 TAQMVTQLAATKPSPIVPKASPK-ALMTPPPPPPGLPRALAAAAKLLGLPSSPVASAMHAK 640

Qy    616 --EEPTPTTP----EEPAPTTP---------KAAAPNTPKE----------PAPTTPKEP 650
Db    641 VTPRPLPASPVPMAASPASLGPDAARVALATNAASPGAKPEAAGGNGTLMAPMGAANTQM 700

Qy    651 APTTPKEPAPTTPKETAPT-------------TPKGTAPT-------------------- 677
Db    701 APIGAAGAAQTAPMGAAHTHVSPMGAGGATQMSPTGAANTHMSPIGAGGATQMSPMGAAN 760

Qy    678 TLKEPAPTTPKKPAPKELAPTTTKEP--------TSTTSDKPAPTTPKG--TAPTTPKEP 727
Db    761 TQMSPMGATTTQMSPMGAAATTQPSPMGAAATQVTATSAGNTMQVSPMGAATPPQTPSVG 820

Qy    728 APTTPKEPAPTTPKGTAPTTLKEP--APTTPKKPAPKELAPTTTKGP----TSTTSDKP- 780
Db    821 AATTP-QPSPM----GAATTLMSPMGAATTPQ---PSPMGAVTTQPPPMAATNTTQPPPM 872

Qy    781 APTTPKET---APTTPKEP---APTTPKKP-APTTPETPP-------PTTSEVSTPTTTK 826
Db    873 AASTPQSTPMGAATTTQSPPMGATTTQSPPMGASTPQAPPTVAGSPTPPPPIPPSPTAQT 932

Qy    827 EPTTIHKSPDESTPEL--SAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTER 884
Db    933 SPQPMSKSPPPDPPKAPSAAAQTSPAAHVANASPGV-TAVSPA---PIGVTEASPSADGA 988

Qy    885 DLRTTPETTTAAPKMT-KETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTL 943
Db    989 RLSPGPTAATDGPKASPAATADVTEAATD--VTAAATAVPA----EAAP---------TK 1033

Qy    944 APKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMP 1003
Db   1034 AKRSSSSSSSSSSSSSSSSSSSSSSSSSDSDSSSSSSESNPASPA----------------P 1078

Qy   1004 RVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAE 1063
Db   1079 AVGDGQQQMTPGAAQSVPP-------VTEAAVQ-------------EAAAAAAAAAGAE 1117

Qy   1064 GE 1065
Db   1118 RE 1119


RESULT   9
T25697
hypothetical protein F16F9.2 - Caenorhabditis elegans
C;Species: Caenorhabditis elegans
C;Date: 15-Oct-1999 #sequence_revision 15-Oct-1999 #text_change 18-Feb-2000
C;Accession: T25697
R;Fulton, B.
submitted to the EMBL Data Library, August 1996
A;Description: The sequence of C. elegans cosmid F16F9.
A;Reference number: Z20071
A;Accession: T25697
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-1229 <FUL>
A;Cross-references: EMBL:U67956; PIDN:AAB07691.1; GSPDB:GN00028; CESP:F16F9.2
A;Experimental source: strain Bristol N2; clone F16F9
C;Genetics:
A;Gene: CESP:F16F9.2
A;Map position: X
A;Introns: 35/1; 361/1; 384/1; 482/2; 517/1; 971/1; 1021/2; 1179/3

  Query Match             9.3%;  Score 680;  DB 2;  Length 1229;
  Best Local Similarity   26.6%;  Pred. No. 7.3e-22;
  Matches  334;  Conservative  105;  Mismatches  394;  Indels  422;  Gaps   60;
```

```
Qy      23 SSQDLSS-CAGRCGEGYSRDA------TCNCDYNCQHYMECCPDFKRVCTAAEVHNPTSP 75
Db     171 TSFDLSTGCNNYDNQGHSQTDFPGFYWNIDCDNNC------------------------ 205

Qy      76 PSSKKAPPPSGASQTIKSTTKRSPKPPNKKKTKKVIESEEITEEHSVSENQESSSSSSSS 135
Db     206 ----------GGT---KSSTTTSSTSESSTTTSSTSESSTTTSSTSESSTTTSSTSESST 252

Qy     136 SSSSTIWKIKSSKNSAANRELQKKLKVKDNKKNRTKKKPTPKPPVVDEAGSGLDNGDFKV 195
Db     253 SSSTT------APATPTTTSCTKEKPTPPTTTSCTKEKPT--PPHHD------------ 291

Qy     196 TTPDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTN 255
Db     292 TTPCT-----KKKTTTSKTCTKKTTTPVPT--PSSSTTESSS------APVPTPSSSTT- 337

Qy     256 KQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPT 315
Db     338 ---------ESSSAPVTSSTTESSS---APV------PTPSSSTTESSSAPVTSSTTESS 379

Qy     316 TPKEPASTTPKEPTPTTIKSAPTTPKEPAPTT-----TKSAP----TTPKEPAPTTTKEP 366
Db     380 SAPVTSSTTESSSAPVPTPSSSTTESSSAPVTSSTTESSSAPVTSSTTESSSAPVTSS-- 437

Qy     367 APTTPKEPAP----TTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPT 422
Db     438 --TTESSSAPVTSSTTESSSAPVPTPSSSTTESSSAPVT----SSTTESSSAPV----PT 487

Qy     423 P---TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPSPT 479
Db     488 PSSSTTESSSAPVT---SSTTESSSAPV-------PTPSSSTTESSSAPAPTPS---SST 534

Qy     480 TPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPT-TPKKP 538
Db     535 TESSSAPVTSS---TTESSSAPVPTPSSSTTESSSTPVTSS----TTESSSAPVPTPSSS 587

Qy     539 APTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTT 598
Db     588 TTESSSAPVPT----PSSSTTESSSAPAPTPSSSTTESSSAPVT----SSTTESSSAP-V 638

Qy     599 PEKPAPTTPEELAPT-TPEEPTPTTPEEPAP-----TTPKAAAPNTPKEPAPTTPKEPAP 652
Db     639 PTPSSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESSSAPVT----SSTTESSSAP 694

Qy     653 TTPKEPAPTTPKETAPT-TPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKP 711
Db     695 VT----SSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESSSAPVPTPS-SSTTESSS 749

Qy     712 APTTPKGTAPTTPKEPAP---TTPKEPAPT-TPKGTAPTTLKEPAPTTPKKPAPKELAPT 767
Db     750 APVTSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESSVAPV 809

Qy     768 TTKGPTSTTSDKPAPTTPKETAPTTPKEPAP-----TTPKKPAPTTPETPPPTTSEVSTP 822
Db     810 PTPSSSSNITSSAPSSTPFSSSTESSSVPVPTPSSSTTESSSAPVSSSTTESSVAPVPTP 869

Qy     823 T-----TTKEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTA 877
Db     870 SSSSNITSSAPSSI---PFSSTTESFSTGT-TVTPSSSKYPGSQTETSVSSTTETTIVPT 925

Qy     878 KDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITAT--TTQVTSTTTQDTTPFKI 935
Db     926 KTTTSVTTPSTTTITTTVCSTGTNSAGETTSGCSPKTVTTTVPTTTTTSVTTSSTTTITT 985

Qy     936 TTLKTTTLAPKVTT---TKKTITTT-EIMNKPEETAKPKDRATNSKATTPKPQKP----- 986
Db     986 TVCSTGTNSAGETTSGCSPKTITTTVPCSTSPSETA--------SESTTTSPTTPVTTVV 1037

Qy     987 ------------TKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAM 1034
Db    1038 STTVVTTEYSTSTKPGGEITTTFVTKNIPTTYLTTIAPTP--SVTTVTNFTPTT-ITTTV 1094

Qy    1035 LQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQG 1094
Db    1095 CSTGT------------NSAGETTSGCSPKTVTTTVPCSTGTGEYTTEATTLVTTAVTT 1141

Qy    1095 IIINPMLSDETNICNGKPVDGLTTLRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGIP 1154
Db    1142 TVVTTESSTGTNSA-GKTTTGYTTKSVPTT------YVTTLAPSAPVTPATN-----AVP 1189

Qy    1155 SPIDTVFTRC----NCEGKT 1170
Db    1190 TTITT--TECSAATNAAGET 1207
```

RESULT 7
S49915
extensin-like protein - maize
C;Species: Zea mays (maize)
C;Date: 05-Mar-1995 #sequence_revision 12-May-1995 #text_change 29-Oct-1999
C;Accession: S49915
R;Rubinstein, A.L.; Broadwater, A.H.; Lowrey, K.; Bedinger, P.A.
submitted to the EMBL Data Library, June 1994
A;Description: Pex genes: pollen-specific genes with extensin-like domains.
A;Reference number: S49915
A;Accession: S49915
A;Status: preliminary
A;Molecule type: DNA
A;Residues: 1-1188 <RUB>
A;Cross-references: EMBL:Z34465; NID:g600117; PIDN:CAA84230.1; PID:g600118

Query Match 10.7%; Score 784; DB 2; Length 1188;
Best Local Similarity 29.9%; Pred. No. 2.8e-26;
Matches 230; Conservative 59; Mismatches 359; Indels 122; Gaps 33;

```
Qy     174 PTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPRPSL----PPN 229
Db     460 PTPHSPPAD---------DYVPPTPPVPGKSPPATSPSPQVQPPAASTPPPSLVKLSPPQ 510

Qy     230 SDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKV 289
Db     511 APVG-----------------------SPPPPVKTTSPPAPIG----SPSPPPPVSVV 541

Qy     290 -----LAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPA 344
Db     542 SPPPPVKSPPPPAPVGSPPPPEKSPPPPAPVASPPPPVKSP--PPPTLVASPPPPVKSPP 599

Qy     345 PTTTKSAPTTP--KEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTP-KEPAPTT 401
Db     600 PPAPVASPPPPVKSPPPPTPVASPPPPAPVASSPPPMKSPPPPTPVSSPPPPEKSPPPPP 659

Qy     402 PKKPAPTTPKEPAPTT------PKE---PTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKP 452
Db     660 PAKSTPPPEEYPTPPTSVKSSPPPEKSLPPPTLIPSPPPQEKPTPPSTPSKP-PSSPEKP 718

Qy     453 APTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPK 512
Db     719 SP--PKEPVSSPPQTP----KSSPPPAPVSSPPPTPVSSPPALAPVSSPPSVKSSP---- 768

Qy     513 EPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAP 572
Db     769 PPAPLSSPPPAPQVKSSPPPVQVSSP-PPAPKSSPPLAPVSSPPQVEKTSPPPAPLSSPP 827

Qy     573 TTPKETAP----TTP----KKLTPTTPEKLAPTTPEKPAP-----TTPEELAPTTPEEPT 619
Db     828 LAPKSSPPHVVVSSPPPVVKSSPPPAPVSSPPLTPKPASPPAHVSSPPEVVKPSTP--PA 885

Qy     620 PTTPEEPAPTTPKAAAPNTPKE-PAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTT 678
Db     886 PTTVISP-PSEPKSSPPPTPVSLPPPIVKSSPPPAMVSSP-PMTPKSSPPPVVVSSPPPT 943

Qy     679 LKE---PA-----PTTPKK---PAPKELAPTTTKE---PTSTTSDKPAPTTPKGTAPTTP 724
Db     944 VKSSPPPAPVSSPPATPKSSPPPAPVNLPPPEVKSSPPPTPVSSPPPA---PKSSPPPAP 1000
```

```
A;Cross-references: EMBL:U48363; NID:g1666688; PID:g1666689; PIDN:AAB18732.1
C;Genetics:
A;Gene: Naca
A;Map position: 10
A;Introns: 24/1; 1996/1; 2024/3; 2050/3; 2099/3; 2142/3; 2183/3
A;Note: differential splicing converts alphaNAC into a tissue-specific DNA-binding activ
C;Keywords: alternative splicing; DNA binding; transcription factor

  Query Match           11.1%;  Score 812;  DB 2;  Length 2187;
  Best Local Similarity  27.4%;  Pred. No. 3.4e-27;
  Matches  311;  Conservative  128;  Mismatches  449;  Indels  246;  Gaps   51;

Qy      71 NPTSPPSSKKAPPPSGASQTIKSTTKRSPKPPNKKKTKKVIESEEITEEHSVSENQESSS 130
           :|| ||||| || ||  :            ||: |    |     |: :   :|   |
Db     746 SPT-PPSSKGAPVPSTGA-----------PPSPKGAPIVPTESSISSKQVPAEILPSPQ 792

Qy     131 SSSSSSSSSTIWKIKSSKNSAANRELQKKLKVKDNKKNRTKKKPTPKPPVVDEAGSGLDN 190
           :   ::|  |  ::| |       :      |      :|     ||
Db     793 KTPEVTASRLISAVQSPKVDPIMSD------VTPTSPKKTSATAVPK------------ 833

Qy     191 GDFKVTTPDTSTTQHNK-----VSTSPKITTAKPINPRPSLPPNSDTSKETSL--TVNKE 243
               ||| |   |       | ||      | |    :|    || : :|     ||
Db     834 --------DTSATLSLKSVPAVTSLSPPKAPVAPSNEATIVPTEIPTSLKNALAAATPKE 885

Qy     244 T------TVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKA 297
           |       | :     | |  |  |    || | |   |  ::||::|:     |  |
Db     886 TLATSIPKVTSPSPQKTPKSVSLKGAPAMTSKKAT---EIAASKDVSPSQ--FPKEVPLL 940

Qy     298 ETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKE-PAPTTTKSAPTTPK 356
           :    |  : || |   |     :: | :  | |:   || ||: | |  :|  | ||
Db     941 QHV---PPTSPPKSPVSDTLSGALTSPPPKGPPATLAETPTYPKKSPKPAASKKTPATPS 997

Qy     357 ---------EPAPTTTKEPAPTTPKEPAPTTTKEPAPTT--TKSAPT------------- 392
                     |  | : | |    ||| : |:: : || |  :|  |:
Db     998 PEGVTAVPLEIPPCSKKAPKTAAPKESSATSSSKRAPKTAVSKEIPSKGVTAVPLEISLP 1057

Qy     393 --------TPKEPAPTTPKK-PAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKE 443
                  || | : ::||: |    |||   | |   |   |:   |  : |   || |
Db    1058 LKETSKSATPGEKSASSPKRSPKTAGPKE---TPPGGVTAVPPEISLPPKETPQNATPNE 1114

Qy     444 PAPTAPKKPAPTT--PKEPAP----TTPKEPAPTTTKEPSPTTPK-----EPAPTTTKSA 492
               : :| :| |  |||  |      | |      | |    ||     | | |    :
Db    1115 SLAASSQKRSPKTSVPKETPPGGVTAMPLEIPSAPQKAPKTAVPKQIPTPEDAVTILAGS 1174

Qy     493 PTTTKEPAPTTT-KSAPTTPK--------EPSPTTTKEPAPTTPKEPAPTTPKKPAPTT- 542
           | : |: : |   | || ||         | ||:  |     |||  : | | | :| |
Db    1175 PLSPKKASKTAAPKEAPATPSVGVIAVSGEISPSPKKTSKTAAPKENSATLPPKRSPKTA 1234

Qy     543 -PKE-----------PAPTTPKEPAP-------TTTKKPAPTAPKEPAPTTPKETAPTT 582
            |||            |:  :|  | |       | | | || |  | :| :||:   |
Db    1235 APKETPATSSEGVTAVPSEISPSPPTPASKGVPVTLTPKGAPNALAE-SPASPKKVPKTA 1293

Qy     583 PKKLTPTTPEKLAPTTPEK-PAPTTPEELAPTTPEEPTPTT--------PEEPAPTTPKA 633
             : | |||      :|:| |    |:| : | | : || |         | |     |
Db    1294 APEETSTTP------SPQKIPKVAGPKEASATPPSKKTPKTAVPKETSAPSEGVTAVPLE 1347

Qy     634 AAPNTPKEPAPTTPKE-PAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTP----- 687
             |:  | |    ||| |||:   | | | | :  |:  ||:     ||  ||||
Db    1348 IPPSPRKAPKTAAPKETPAPS--PEGATTAPVQIPPSPRKGSKKAGSKE-TPTTPSPEGV 1404

Qy     688 --------------KKPAPKELAPTTTKEPTS------TTSDKPAPTTPKGTAPTTPKE 726
                         |  :|||   | : :  |       || : |   |    |:  |
Db    1405 TAAPLEIPISSKKTSKMASPKETLVTPSSKKLSQTVGPKETSLEGATAVPLEIPPSHKKA 1464

Qy     727 PAPTTPKE-PAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTP 785
           |    ||: |   :||  ||||| | :|:::||| |||  || : :  |:    :||| |
Db    1465 PKTVDPKQVPLTPSPK-DAPTTLAE-SPSSPKK-APKTAAPPSER-VTTVPPEKPA-TPQ 1519

Qy     786 KETAPTTPKEPAPTTPKKPAPTTPETP-----PPTTSEVSTPTTTK------EPTTIHKS 834
           | :  |  | | |   :: | :: |||      ||  :  |   |:|       | |:  |
Db    1520 KASGTTASKVPVPAETQEVAVSSRETPVTPAVPPVKNPSSHKKTSKTIELKEAPATLPPS 1579

Qy     835 PDESTPELSAEPTPKALENSPKE-PGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETT 893
           | :|    |::  |:   ::|||| |  |: | |  |   :  ||     :    | |:
Db    1580 PTKSPKIPSSKKAPRT--SAPKEFPASPSIK-PVTT--SLAQTAPPSLQKAPSTTIPKEN 1634

Qy     894 TAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKT 953
           |||  :      : |:  :   |: :   :|    | | : ||: :   |    |  :|
Db    1635 LAAPAV----LPVSSKSPAAPARASASLSPATAAPQTAPKEATTIPSCKKAAATETPIET 1690

Qy     954 ITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPT 1013
           |    :   |:||::     : ||     |       ||| :|:|:  |:|
Db    1691 STAPSLEGAPKETSE----TSVSKVLMSSP------PKKASSSKRASTLP---------- 1730

Qy    1014 PRKMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETP 1067
               :|:| |   | ::         |  | : |   ::| | ||      ||
Db    1731 ----ATTLPSLKEASVLS---------PTATSSGKDSHISPVS-DACSTGTTTP 1770


RESULT   6
S48478
glucan 1,4-alpha-glucosidase (EC 3.2.1.3) - yeast (Saccharomyces cerevisiae)
N;Alternate names: extracellular glucoamylase; mucin-like protein MUC1; protein YIR01
C;Species: Saccharomyces cerevisiae
C;Date: 10-Sep-1999 #sequence_revision 10-Sep-1999 #text_change 12-Nov-1999
C;Accession: S48478; A26877; B26877; S27281; JC6123
R;Rowley, K.
submitted to the EMBL Data Library, October 1994
A;Reference number: S48478
A;Accession: S48478
A;Molecule type: DNA
A;Residues: 1-1367 <ROW>
A;Cross-references: GB:Z47047; EMBL:Z38061; NID:g603997; PID:g763364; GSPDB:GN00009;
R;Yamashita, I.; Nakamura, M.; Fukui, S.
J. Bacteriol. 169, 2142-2149, 1987
A;Title: Gene fusion is a possible mechanism underlying the evolution of STA1.
A;Reference number: A91831; MUID:87194600
A;Accession: A26877
A;Molecule type: DNA
A;Residues: 1-242 <YAM>
A;Cross-references: EMBL:M16164; NID:g172522; PIDN:AAA35014.1; PID:g172525
A;Accession: B26877
A;Molecule type: DNA
A;Residues: 762-1331 <YA2>
A;Cross-references: EMBL:M16165; NID:g172523; PIDN:AAA35015.1; PID:g172526
R;Pardo, J.M.; Ianez, E.; Zalacain, M.; Claros, M.G.; Jimenez, A.
FEBS Lett. 239, 179-184, 1988
A;Title: Similar short elements in the 5' regions of the STA2 and SGA genes from Sacc
A;Reference number: S27281; MUID:89031230
A;Accession: S27281
A;Molecule type: DNA
A;Residues: 1-31 <PAR>
A;Cross-references: EMBL:X13857; NID:g4551; PIDN:CAA32069.1; PID:g4552
R;Lambrechts, M.G.; Bauer, F.F.; Marmur, J.; Pretorius, I.S.
Proc. Natl. Acad. Sci. U.S.A. 93, 8419-8424, 1996
A;Title: Muc1, a mucin-like protein that is regulated by Mss10, is critical for pseud
A;Reference number: JC6123; MUID:96323237
A;Accession: JC6123
A;Status: nucleic acid sequence not shown
A;Molecule type: DNA
A;Residues: 1-1367 <LAM>
A;Cross-references: GB:U30626; NID:g1304386; PIDN:AAC49609.1; PID:g1304387
C;Genetics:
A;Gene: SGD:MUC1; STA2; MAL5; DEX2; SGD:S0001458
A;Cross-references: MIPS:YIR019c; SGD:S0001458
A;Map position: 9R
C;Superfamily: yeast glucan 1,4-alpha-glucosidase homolog; glucan 1,4-alpha-glucosida
C;Keywords: glycosidase; hydrolase; polysaccharide degradation; transmembrane protein
F;5-21/Domain: transmembrane #status predicted <TM1>
F;1350-1366/Domain: transmembrane #status predicted <TM2>

  Query Match           11.1%;  Score 810.5;  DB 1;  Length 1367;
  Best Local Similarity  27.2%;  Pred. No. 2.5e-27;
  Matches  332;  Conservative  114;  Mismatches  519;  Indels  255;  Gaps   54;
```

```
Db   1048 ASTEETTYAPTEETTYAPAEETPYEPTEETTYAPTEETTYAPTEETTYAPTEETTYAPTE 1107

Qy    783 TTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPT--TIHKSPDESTP 840

Db   1108 ETTYAPAEETPYEPTEETTYAPTEETTYAPTEETMYAPIEETTYGPTEETTY-APTEATT 1166

Qy    841 ELSAEPTPKA-LENSPKEPG-----VPTTKTPAATKPEMTTTAKDKTTERDLRTTP---- 890

Db   1167 YAPTEETPYAPTEETTYEPTGETTYAPTEETTYAPTEETTYAPTEETTYAPTEETPYEPT 1226

Qy    891 ETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTT----------PFKITT--- 937

Db   1227 EETTYAP-----TEETTYEPTEETTYAPTEETTYAPTEETTYAPTEETMYAPIDETTYGP 1281

Qy    938 LKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRAT--NSKATTPKP-QKPTKAPKKPT 994

Db   1282 TEETTYAPTEATTYAPTEETPYAPTEETTYEPTGETTYAPTEETTYAPTEETTYAPMEET 1341

Qy    995 -----------STKKP-KTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRP 1041

Db   1342 PYEPAEESTSTVSTEKPCNTEEFTDEPTDEPTDE--PSDEPTDEPTDEPTDLPTDEPSTP 1399

Qy   1042 --NQTPNSKLVEVNPKSEDAG 1060

Db   1400 CDNQGINGIGVENKVRYNNAG 1420
```

RESULT 4
T16251
hypothetical protein F35A5.1 - Caenorhabditis elegans
C;Species: Caenorhabditis elegans
C;Date: 20-Sep-1999 #sequence_revision 20-Sep-1999 #text_change 18-Feb-2000
C;Accession: T16251
R;Leimbach, D.
submitted to the EMBL Data Library, January 1996
A;Description: The sequence of C. elegans cosmid F35A5.
A;Reference number: Z18485
A;Accession: T16251
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-1274 <LEI>
A;Cross-references: EMBL:U46675; NID:g1166613; PID:g1166621; PIDN:AAB52641.1; GSPDB:GN0C
A;Experimental source: strain Bristol N2; clone F35A5
C;Genetics:
A;Gene: CESP:F35A5.1
A;Map position: X
A;Introns: 1272/2

Query Match 11.6%; Score 846; DB 2; Length 1274;
Best Local Similarity 27.6%; Pred. No. 7.2e-29;
Matches 313; Conservative 119; Mismatches 455; Indels 248; Gaps 60;

```
Qy     76 PSSKKAPPPSGASQTIKSTTKRSP----KPP---NKKKTKKVIESEEITEEHSVSENQES 128

Db    185 PSPKKAAPSKEHDPIVPPTPIKNPAKKWKPPWEDDEVPTEEIKEPEPATRKVPALKKKEP 244

Qy    129 SSSSSSSSSSSTIWKIKSSKNSAANRELQKKLKVKDNK-KNRTKK-KP-----TP----- 176

Db    245 STSVKPVSDPSPTKKV----------PVKKEPEVPPTPIKNPTKKWKPPWEDETPVEEVK 294

Qy    177 KPPVVDEAGSGLDNGD----FKVTTPDTSTTQHNKVSTSPKITTAKPI-NPRPSLPPNSD 231

Db    295 EPPVPEKKAPVLKKKDPAPAAKARDPSPSKAAPKKVEPSSPVVPPTPVKNPVKKYKPPWE 354

Qy    232 TSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLA 291

Db    355 VDDEPAEEVKKPSAPEKKTPVLKRKEPEPSSTTPSSDPSPKKAAPAVKPRDSSPKKATPL 414

Qy    292 KPTPKAE----TTTKGPA----------------LTTPKEPTPTTP----KEPASTTPK 326

Db    415 QADPKAQEVPPTPVKNPVKKYKPPWEVDDEDPVEEVKQPEAPAKKTPVLKRKEPAAKDTA 474

Qy    327 EP----TP-TTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPA---------- 367

Db    475 KPATSKTPETPEKKDPVKPRDSSPKKVAAKPDSAQAPA-TPVKNPVKKWRPPWEDDETPA 533

Qy    368 -----PTTPKEPAPTTTKEPAPT-------TTKSAPTTPKEP-----APTTPKKPAPTTP 410

Db    534 DDVSKPTDAKKTPSLAKKDPAPAKESLKPKADTKAPAKPRDPSPKKVAPTAPEKKTPVLA 593

Qy    411 KE----PAPTTPKEPTPTTPKEPAPTTKEPAPTTPK-EPAPTAPKKPAP------TTPKE 459

Db    594 KKEPAGPADSKTKEPEKSKPRDPSPKKAVPAKPVPKTEVAPAAVKKPEPISKPKDTAPKK 653

Qy    460 PAPTTPKEPAPTTTKEP----SPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTP-KEP 514

Db    654 AEPNSPVVP-PTPVKNPVKKWKPPWEDDDAPAKPVSLPEPEKK-TPVLAKKAPTKPDSEA 711

Qy    515 SPTTTKEPAPTTPK--EPAPTTPKKPAP--TTPKEPAPTTPKEPAPTTTKKPAPTA-PKE 569

Db    712 AADPVSGPSSKDPKLAKKAPVKPRDPSPMKAVPIKPAPKT--EVPPAVVKKPEPVAKSRD 769

Qy    570 PAPTTPKETAPTTPKKLTPTTP-----EKLAPTTPEKPAPTTPEELAPTTPEEP-----T 619

Db    770 PSPKKAK-AEPNSP--VVPPTPVKNPVKKWKPPWEDDDAPAEPVNVPEPEKKTPVLAKKT 826

Qy    620 PTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTT-PKEPAPTTPKETAPTTPKGTAPTT 678

Db    827 PVKPRDPSPKKAVPAKPSTKTDAPPVSVKKPEPVSKPKEPSPKKAEPNSPVVP----PTP 882

Qy    679 LKEPAPTTPKKPAPKELAPT-TTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAP 737

Db    883 VKNPVKKW-KPPWEDDDEPTEEVKKPSE--PEKKTPVLAK-KEPEKPKD-APKVAAKPRD 937

Qy    738 TTPKGTAPTTLKEPAPT-------TPKKPAP-----KELAPTTTKGPTS------TTSDK 779

Db    938 PSPKKAVPE--KEPAKVAAKPRDLSPKKAIPIPANTQEAPPTPVKNPVKKWKPPWEDDDE 995

Qy    780 P-----APTTPKET------APTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEP 828

Db    996 PAEPVSAPEPEKKTPVLAKKAPAKPRDP---SPKKAAPVAAK-PDPKIPEV-PPTPVKNP 1050

Qy    829 TTIHKSP----DESTPELSA-EP---TPKALENSPKEPGV-------------PTTKTPA 867

Db   1051 VKKWKPPWEDDDEPSEPVSAPEPEKKTPVLAKKAPTKPATKPDSEAAADPVSGPTSKDPK 1110

Qy    868 AT------KPEMTTTAKDKTTERDLRTTPETT--TAAPKMTKETATTTEKTTESKITATT 919

Db   1111 LSKKAPVEKPKPTTDPKDDKLKPSPAKKPEKAPEPAAPKKWKPVWDDDPDEPEADFTVPA 1170

Qy    920 TQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKAT 979

Db   1171 PSKKPDTEDPADPLG---------GPKTKDPK--------LNKKAPAEKPTEK------- 1206

Qy    980 TPKPQKPTKAPKKPTSTKKPKTMPRVRKPK------------TTPTPRKMTSTMP 1022

Db   1207 -PKPKEVSKEPPKPTEPPKP-AAPKKWKPPWEDDPDEPEADFTMPAPKKPDTEDP 1259
```

RESULT 5
T30826
nascent polypeptide-associated complex alpha chain, muscle splice form gp220 - mouse
N;Alternate names: alpha-NAC protein
C;Species: Mus musculus (house mouse)
C;Date: 22-Oct-1999 #sequence_revision 22-Oct-1999 #text_change 24-Nov-1999
C;Accession: T30826
R;Yotov, W.V.; St-Arnaud, R.
Genes Dev. 10, 1763-1772, 1996
A;Title: Differential splicing-in of a proline-rich exon converts alphaNAC into a mus
A;Reference number: Z20889; MUID:96312450
A;Accession: T30826
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-2187 <YOT>

Matches 312; Conservative 104; Mismatches 361; Indels 230; Gaps 55;

```
Qy    289 VLAKPTP-KAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTT 347
Db    758 VVIQPAPIKAASDEPIPTDTPSDEPTPS----------DEPTPS---DEPTPSDEPTPSD 804

Qy    348 TKSAPTTPKEPAPTTTKEPAPTTPKEPA----PTTTKEPAPTTTKSAPTTPKEPAPTTPK 403
Db    805 EPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTP 864

Qy    404 KPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPT-TPKEPAPTAPKKPAPTTPKEPAP 462
Db    865 SDEPTPSDEPTPS--DEPTPS--DEPTP-SDEPTPSETPEEPIPTDTPSDEPTPSDEPTP 919

Qy    463 TTPKEPA--PTTTKEPSPT-TPKEPAPTTTKS-APTTTKEPAPT---TTKSAPTTPKEPS 515
Db    920 SDEPTPSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPT 979

Qy    516 ----PTTTKEPAPT-TPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEP 570
Db    980 PSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSDEP 1039

Qy    571 APT---TPKETAPTTPKKLTPTTPEKLAPTTPEKPAPT---TPEELAPTTPEEPTPT-TP 623
Db   1040 TPSDEPTPSE----TPEEPIPTDTPSDEPTPSDEPTPSDEPTPSD-EPTPSDEPTPSETP 1094

Qy    624 EEPAPTTPKAAAPNTPKEPAPTTPKEPA--PTTPKEPAPT-TPKETAPTTPKGTAPTTLK 680
Db   1095 EEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSD 1154

Qy    681 EPAPTTPKKPAPKELAPTTTKEPT-STTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTT 739
Db   1155 EPTPS--DEPTPSD-EPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPS--DEPTPS- 1208

Qy    740 PKGTAPTTLKEPAPT-TPKKPAPKELAPTTTKGPTSTTSDKPAPT---TPKETAPTTPKE 795
Db   1209 ---DEPTPSDEPTPSETPEEPI----------PTDTPSDEPTPSDEPTPSD-EPTPSDE 1253

Qy    796 PAPTTPKKPAPT-TPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENS 854
Db   1254 PTPS--DEPTPSETPEEPIPTDTPSDEPTPSDEPT----PSDEPTP--SDEPTPSDEPTP 1305

Qy    855 PKEPGVPTTKTPAAT--KPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTE 912
Db   1306 SDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSD----EPTPSDEPTPSDEPTPSDEPTPS 1361

Qy    913 SKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDR 972
Db   1362 DEPTPSDEPTPSETPEEPTP-------TTTPTPTPSTT-----------------PTSG 1396

Qy    973 ATNSKAT--------TPKPQKPTKAP-KKPTSTKKPKTMPRVRKPKTTPTPRKM------ 1017
Db   1397 SGGSGGSGGGGGGGGGTVPTSPTPTPTSKPTSTPAP---TEIEEPTPSDVPGAIGGEHRA 1453

Qy   1018 -------TSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHM- 1069
Db   1454 YLRGYPDGSFRPERNITRAEAAVIF----------AKLL----GADESYGAQSASPYSD 1498

Qy   1070 LLRPH--------------------VFMPE--------VTPDMDYLPRVPNQGI----- 1095
Db   1499 LADTHWAAWAIKFATSQGLFKGYPDGTFKPDQNITRAEFATVVLHFLTKVKGQEIMSKLA 1558

Qy   1096 ---IINPMLSDETNICNG----KPVDGLTTL------RNGTLVAFRGHYFWMLSPFSPPS 1142
Db   1559 TIDISNPKFDD----CVGHWAQEFIEKLTSLGYISGYPDGT-------------FKPQN 1600

Qy   1143 PARRITEVWGIPSPIDTVFTRCNCEGKTFFFKD--SQYWRFTNDIKD 1187
Db   1601 YIKRSESV----ALINRALERGPLNGAPKLFPDVNESYWAF-GDIMD 1642
```

RESULT 3
T31108

cyst germination specific acidic repeat protein precursor - Phytophthora infestans
C;Species: Phytophthora infestans (potato late blight agent)
C;Date: 22-Oct-1999 #sequence_revision 22-Oct-1999 #text_change 22-Oct-1999
C;Accession: T31108
R;Goernhardt, B.
submitted to the EMBL Data Library, April 1998
A;Reference number: Z20986
A;Accession: T31108
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-1489 <GOE>
A;Cross-references: EMBL:AF061185; NID:g3851513; PID:g3851514; PIDN:AAC72308.1
C;Genetics:
A;Gene: car90

Query Match 11.8%; Score 863.5; DB 2; Length 1489;
Best Local Similarity 31.7%; Pred. No. 1.5e-29;
Matches 368; Conservative 55; Mismatches 509; Indels 229; Gaps 51;

```
Qy    107 TKKVIESEEITEEHSVSENQESSSSSSSSSSSSTIWKIKSSKNSAANRELQKKLKVKDNK 166
Db    282 TPTVGYSTEETEGQHVTGGYEPSDETEAPTEGTTY--VPREETTAAPSE-----DTTYAP 334

Qy    167 KNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTT------QHNKVSTSPKITTAKPI 220
Db    335 REVTPYAPTEKPYDVEETTYVTEESTYAPTKSETNAPTERMHYAHIEKPCDTEVTMYAPT 394

Qy    221 NPRPSLP--------------PNSDT----SKETSLTVNKETTVETKETTT---TNKQT 258
Db    395 EETTYAPTEETTYAPTEETTYAPTEETPYEPTEETTYTPTEETTYAPTEETTYAPTEKTT 454

Qy    259 STDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPK 318
Db    455 YAPTEETTYAPTEETPYEPTEETTYAPTKETTYAPT---EETTYASTEETTYAPTEETTY 511

Qy    319 EPASTTPKEPTPTTIKSAPTTPKEPAPT-TTKSAPTTPKEPAPT--TTKEPAPTTPKEPA 375
Db    512 APAEETPYEPTEET-TYAPTEETTYAPTEETTYAPTEETTYAPTEETTYAPAEETPYEPT 570

Qy    376 PTTTKEPAPTTTKSAPT----------TPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTT 425
Db    571 EETTYAPTEETT-YAPTEETMYAPIEETTYAPTEETTYAPAEETPYEPTEETTYAPTEET 629

Qy    426 PKEP--------------AP---TTKEPAPTTPKEP------APTAPKKPAPTTPKEPAP 462
Db    630 TYAPTEETTYASTEETTYAPTEETTYAPAEETPYEPTEETTYAPTEETTYAPTEETTYAP 689

Qy    463 TTPKEPAPT--TTKEPSPTTPKEP------APT-TTKSAPT----------TTKEP---- 499
Db    690 TEETTYAPTEETTYAPAEETPYEPTEETTYAPTEETTYAPTEETMYAPIEETTYGPTEET 749

Qy    500 --APT-TTKSAPT--TPKEPSPTTTKEP------APTTPKEPAPTTPKKPAPTTPKEPAP 548
Db    750 TYAPTEATTYAPTEETPYAPTEETTYEPTGETTYAPTEETTYAPTEETTYAPTEETTYAP 809

Qy    549 T--TPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTT 606
Db    810 TEETPYEPTEETTYAPTEETPYEPTEET--TYTPTEETTYAPTEETTYAPTEKTTYAPTE 867

Qy    607 PEELAPT--TPEEPTPTTPEEP------APT----------TPKAAAPNTPKEPAPTTPK 648
Db    868 ETTYAPTEETPYEPTEETTYAPTKETTYAPTEETTYASTEETTYAPTEETTYAPAEETPY 927

Qy    649 EP------APTTPKEPAPTTPKETAPTTPKGTAPT--TLKEPAPTTPKKPAPK-ELAPT- 698
Db    928 EPTEETTYAPTEETTYAPTEETTYAPTEETTYAPTEETTYAPAEETPYEPTEETTYAPTE 987

Qy    699 -TTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEP------APTTPKGTAPT----- 746
Db    988 ETTYAPTEETMYAPIEETTYAPTEETTYAPAEETPYEPTEETTYAPTEETTYAPTEETTY 1047

Qy    747 ------------TLKEPAPTTPKKP------APKE---LAPT--TTKGPTSTTSDKPAP 782
```

J. Clin. Invest. 87, 77-82, 1991
A;Title: Human bronchus and intestine express the same mucin gene.
A;Reference number: A61257; MUID:91086481
A;Accession: A61257
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 'T',1925-1948,'TTS',1952-1954 <JAN>
A;Experimental source: bronchus
R;Xu, G.; Huan, L.; Khatri, I.; Sajjan, U.S.; McCool, D.; Wang, D.; Jones, C.; Forstner,
Biochem. Biophys. Res. Commun. 183, 821-828, 1992
A;Title: Human intestinal mucin-like protein (MLP) is homologous with rat MLP in the C-t
A;Reference number: PQ0328; MUID:92198477
A;Accession: PQ0328
A;Molecule type: mRNA
A;Residues: 2328-2468 <XUG>
A;Cross-references: GB:M86523
A;Experimental source: small intestine
A;Accession: PQ0329
A;Molecule type: protein
A;Residues: 2328-2342,'K',2344-2354 <XUG1>
C;Genetics:
A;Gene: GDB:MUC2
A;Cross-references: GDB:120203; OMIM:158370
A;Map position: 11p15.5-11p15.5
C;Superfamily: von Willebrand factor; von Willebrand factor type A repeat homology; von
C;Keywords: glycoprotein; intestine; tandem repeat
F;2766-2834/Domain: von Willebrand factor type C repeat homology <VWC>

Query Match 14.7%; Score 1073; DB 2; Length 3020;
Best Local Similarity 27.1%; Pred. No. 4.3e-38;
Matches 374; Conservative 101; Mismatches 502; Indels 402; Gaps 50;

```
Qy      29 SCAGRCGEGYSRDATCNCDYNCQHYMEC-------------CPDFKRVCTAAEVHNPTS 74
           ||:  |  |   |  | |     :  ||               || |       : :||
Db    1079 SCS--CDTG--GDCECFCSAVASYAQECTKEGACVFWRTPDLCPIF------CDYYNPPH 1128

Qy      75 PPSSKKAPPPSGASQTIKSTT--------------KRSPK--PPNKKKTKKVIESEE-- 115
             |  : : :| ::                   | ||  |  ::  || : :::
Db    1129 ECEWHYEPCGNRSFETCRTINGIHSNISVSYLEGCYPRCPKDRPIYEEDLKKCVTADKCG 1188

Qy     116 --ITEEH----SVSENQESSSSSSSSSSSSTIWKIKSSKNSAANRELQKKLKVKDNKKNR 169
             : : |    :    :|:  |   ::||  : : :  |                    |:
Db    1189 CYVEDTHYPPGASVPTEETCKSCVCTNSSQVVCRPEEGK----------------ILNQ 1231

Qy     170 TKKKPTPKPPVVDEAGSGLDNGDF---KVTTPDTSTTQH---NKVSTSPKITTAKPINPR 223
           |:                   :| |   ::  |: :  :|     ::| |   |
Db    1232 TQ-----------------DGAFCYWEICGPNGTVEKHFNICSITTRPSTLTTFTTITL 1273

Qy     224 PSLPPNSDTSKETSLTVNKETTVETKET--TTTNK--------------QTSTDGKEKT 266
           |: |        || |     ||  |  |  :|| |               | ||  :
Db    1274 PTTP--------TSFTTTTTTTTPTSSTVLSTTPKLCCLWSDWINEDHPSSGSDDGDREP 1325

Qy     267 TS--AKETQSIEKTSAKD---------------------------------------- 282
             : ||  | ||
Db    1326 FDGVCGAPEDIECRSVKDPHLSLEQHGQKVQCDVSVGFICKNEDQFGNGPFGLCYDYKIR 1385

Qy     283 ----------LAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTT 332
                    :  |      |:|   |||  |  |||  || ||    | :|||  |  ||
Db    1386 VNCCWPMDKCITTPSPPTTTPSPPPTTTTTLPPTTTPSPPTTTTTTPPPTTTPSPPITTT 1445

Qy     333 IKSAP-TTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAP 391
           | |||  |  |||   |||   | ||||  |  |||  |  |||  | ||||  | |
Db    1446 TTPLPTTTPSPPISTTTTPPPTTTPSP-PTTTPSPPTTTPSPPT-TTTTTPPPTTTPSPP 1503

Qy     392 -TTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPK 450
           |||  |  :|   |  |||  |  ||   |  |||  |  |   | |||:     ||
Db    1504 MTTPITPPASTTTLPPTTTPSPPTTTTTTPPPTTTPSPPTTTPITP-PTSTTTLPPTTTP 1562

Qy     451 KPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAP-TTTKEPAPTTTKSAP- 508
           | |||   | |||   |  |||  |   |   | ||||  | |  |||  | ||||  | |
Db    1563 SPPPTTTTTPPPTTTPSPPTTTTPSPPTITTTTPPPTTTPSPPTTTTTTPPPTTTPSPPT 1622
```

```
Qy     509 TTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPK 568
           |||  |  :||  |  |||  |  ||   |  |||  |  |||  |  |||  |  | |
Db    1623 TTPITPPTSTTTLPPTTTPSPPPTTTTTPPPTTTPSPPTTTTPSPPITTTTTPPPTTTPS 1682

Qy     569 EPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAP 628
             |  |||  : |||     | |||      |||  |  ||    : |||  | |||    |:|
Db    1683 SPITTTP--SPPTT----TMTTPS--PTTTPSSPITTTTTPSSTTTP-SPPPTTMTTPSP 1733

Qy     629 TTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTP-----KETAPTTP----------- 671
           ||  :    |     |||   |  |||  |: | |        | ||||
Db    1734 TTTPSPPTTTMTTLPPTTTSSPLTTTPLPPSITPPTFSPFSTTTPTTPCVPLCNWTGWLD 1793

Qy     672 KG---------------------------------------------------------- 673
           |
Db    1794 SGKPNFHKPGGDTELIGDVCGPGWAANISCRATMYPDVPIGQLGQTVVCDVSVGLICKNE 1853

Qy     674 -------------------------TAPTTLKEPAPTTPKKPAPKELAPTT--TKEPTS 705
                                     | |||:     || : | |  :  ||  |  ||
Db    1854 DQKPGGVIPMAFCLNYEINVQCCECVTQPTTM---TTTTTENPTPTPITTTTTVTPTPTP 1910

Qy     706 TTSDKP----APT-TPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPA 760
           |::  |    ||| ||  |  |    | ||  : |  |||  |  |    | ||  : |
Db    1911 TSTQSPNGLQAPTPTPISTTTTVTPTPTPTGTQTPT-TTPITTTTTVTPTPTPTGTQTPT 1969

Qy     761 PKELAPTTTKGPTST-TSDKPAPTTPKETAPTTPKEPAPTTPKKP-------APTTPETP 812
             :  |||  || | || |     ||  |  |     | ||  : |          |  ||
Db    1970 TVLITTTTTMTPTPTPTSTKSTTVTPITTTTTVTATPTPTGTQTPTMIPISTTTTVTPTP 2029

Qy     813 PPTTSEVSTPTTTK-----EPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPA 867
           |||      || |       | || :  |: |||: |    :     |::      :|
Db    2030 TPTTGSTGPPTHTSTAPIAELTTSNPPPESSTPQTSRSTSSPLTESTTLLSTLPPAIEMT 2089

Qy     868 ATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTT 927
           :| |  | ||   |:      : | :|| :|  |     |||     : |  | :| | |:|:
Db    2090 STAPPSTPTAPTTTSGGHTLSPPPSTTTSPPGTPTRGTTT--GSSSAPTPSTVQTTTTSA 2147

Qy     928 QDTTPFKITT---LKTTTLAPKVTTTKKTITTTEIMNKPEE------------------ 965
             ||  ::|   ::|| | |  ::        :  | |
Db    2148 WTPTPTPLSTPSIIRTTGLRPYPSSVLICCVLNDTYYAPGEEVYNGTYGDTCYFVNCSLS 2207

Qy     966 ----------TAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTPR 1015
                     : |    | || :|| | ||:  | |||   ||   |     |:
Db    2208 CTLEFYNWSCPSTPSPTPTPSK-STPTPSKPSSTPSKPTPGTKPPECPDFDPPR------ 2260

Qy    1016 KMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNP-------------KSEDAGG- 1061
           : | |  : :  :  | : | |     ||  |                 : ||  |
Db    2261 -------QENETWWLCDCFM-ATCKYNNTVEIVKVECEPPPMPTCSNGLQPVRVEDPDGC 2312

Qy    1062 -----------AEGETPHMLL-----------RPHVFMPEVTPDMDYLPRVPNQGIIIN 1098
                       |: || :              :| : |::| :|      | |: |:
Db    2313 CWHWECDCYCTGWGD-PHYVTFDGLYYSYQGNCTYVLVEEISPSVD------NFGVYID 2364
```

RESULT 2
T18262
S-layer protein - Clostridium thermocellum
C;Species: Clostridium thermocellum
C;Date: 15-Oct-1999 #sequence_revision 15-Oct-1999 #text_change 15-Oct-1999
C;Accession: T18262
R;Fujino, T.; Beguin, P.; Aubert, J.P.
J. Bacteriol. 175, 1891-1899, 1993
A;Title: Organization of a Clostridium thermocellum gene cluster encoding the cellulo
e.
A;Reference number: Z18847; MUID:93209931
A;Accession: T18262
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-1664 <FUJ>
A;Cross-references: EMBL:X67506; NID:g296879; PID:g296881; PIDN:CAA47841.1

Query Match 13.0%; Score 950; DB 2; Length 1664;
Best Local Similarity 31.0%; Pred. No. 3.8e-33;

GenCore version 4.5
Copyright (c) 1993 - 2000 Compugen Ltd.

OM protein - protein search, using sw model

Run on: April 26, 2002, 16:31:52 ; Search time 114.61 Seconds
(without alignments)
907.900 Million cell updates/sec

Title: AA6
Perfect score: 7294
Sequence: 1 MAWKTLPIYLLLLLSVFVIQ.....ARAITTRSGQTLSKVWYNCP 1366

Scoring table: BLOSUM62
Gapop 10.0 , Gapext 0.5

Searched: 219241 seqs, 76174552 residues

Total number of hits satisfying chosen parameters: 219241

Minimum DB seq length: 0
Maximum DB seq length: 2000000000

Post-processing: Minimum Match 0%
Maximum Match 100%
Listing first 45 summaries

Database : PIR_68:*
1: pir1:*
2: pir2:*
3: pir3:*
4: pir4:*

Pred. No. is the number of results predicted by chance to have a score greater than or equal to the score of the result being printed, and is derived by analysis of the total score distribution.

SUMMARIES

Result No.	Score	% Query Match	Length	DB	ID	Description
1	1073	14.7	3020	2	A43932	mucin 2 precursor,
2	950	13.0	1664	2	T18262	S-layer protein -
3	863.5	11.8	1489	2	T31108	cyst germination s
4	846	11.6	1274	2	T16251	hypothetical prote
5	812	11.1	2187	2	T30826	nascent polypeptid
6	810.5	11.1	1367	1	S48478	glucan 1,4-alpha-g
7	784	10.7	1188	2	S49915	extensin-like prot
8	695	9.5	1151	2	T18535	high molecular mas
9	680	9.3	1229	2	T25697	hypothetical prote
10	672	9.2	1344	1	A35175	mucin 1 precursor,
11	661.5	9.1	7962	2	I38346	elastic titin - hu
12	649	8.9	3570	2	T45025	mucin MUC5B, trach
13	647.5	8.9	3507	2	T34513	hypothetical prote
14	643.5	8.8	990	2	I51618	nucleolar phosphop
15	633	8.7	489	2	T11622	extensin class 1 p
16	632	8.7	761	2	C84672	hypothetical prote
17	629	8.6	6642	2	T29757	protein UNC-89 - C
18	627.5	8.6	971	2	T19431	hypothetical prote
19	622.5	8.5	839	2	F75518	hypothetical prote
20	607.5	8.3	801	2	T29018	hypothetical prote
21	607.5	8.3	924	2	S27923	gene LF3 protein -
22	605	8.3	379	2	S50125	larval glue protei
23	592.5	8.1	2232	2	T34434	hypothetical prote
24	568.5	7.8	1630	2	A53577	ascites sialoglyco
25	559.5	7.7	350	2	S22456	hydroxyproline-ric
26	559	7.7	875	2	S23760	polyphenolic adhes
27	556.5	7.6	856	2	T16543	hypothetical prote
28	551	7.6	620	2	S06733	hydroxyproline-ric
29	551	7.6	1087	1	QFMSH	neurofilament trip
30	545	7.5	873	2	A47283	calphotin - fruit
31	542	7.4	369	2	S20500	hydroxyproline-ric
32	538.5	7.4	416	2	JU0465	extensin precursor
33	537.5	7.4	1459	2	T32271	hypothetical prote
34	537	7.4	1072	1	A37221	neurofilament trip
35	532	7.3	756	2	T27642	hypothetical prote
36	530.5	7.3	813	2	S70795	vsaA protein precu
37	530.5	7.3	1162	2	JH0557	exo-alpha-sialidas
38	530	7.3	865	2	A47282	calcium-binding pr
39	526	7.2	3534	2	T42567	tegument protein 2
40	522	7.2	328	2	JQ0985	hydroxyproline-ric
41	521.5	7.1	854	2	S02003	neurofilament trip
42	518	7.1	866	2	T45462	membrane glycoprot
43	518	7.1	1611	2	T38236	hypothetical prote
44	513	7.0	1832	2	T31113	mucin-like glycopr
45	512.5	7.0	867	2	T45463	membrane glycoprot

ALIGNMENTS

RESULT 1
A43932
mucin 2 precursor, intestinal - human (fragments)
N;Alternate names: mucin SMUC-41
C;Species: Homo sapiens (man)
C;Date: 10-Mar-1993 #sequence_revision 12-Apr-1996 #text_change 05-Nov-1999
C;Accession: A49963; A45106; B45106; A43932; B33532; A61257; PQ0328; PQ0329
R;Gum Jr., J.R.; Hicks, J.W.; Toribara, N.W.; Siddiki, B.; Kim, Y.S.
J. Biol. Chem. 269, 2440-2446, 1994
A;Title: Molecular cloning of human intestinal mucin (MUC2) cDNA. Identification of t
A;Reference number: A49963; MUID:94132002
A;Accession: A49963
A;Molecule type: mRNA
A;Residues: 1-639 <GU1>
A;Cross-references: GB:L21998
R;Gum Jr., J.R.; Hicks, J.W.; Toribara, N.W.; Rothe, E.M.; Lagace, R.E.; Kim, Y.S.
J. Biol. Chem. 267, 21375-21383, 1992
A;Title: The human MUC2 intestinal mucin has cysteine-rich subdomains located both up
A;Reference number: A45106; MUID:93016075
A;Accession: A45106
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 626-1895 <GU2>
A;Cross-references: GB:M94131; NID:g186395; PIDN:AAA59163.1; PID:g186396
A;Note: sequence extracted from NCBI backbone (NCBIP:116706)
A;Accession: B45106
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 2037-3020 <GU3>
A;Cross-references: GB:M94132; NID:g186397; PIDN:AAA59164.1; PID:g186398
A;Experimental source: colon
A;Note: sequence extracted from NCBI backbone (NCBIP:116698)
R;Toribara, N.W.; Gum Jr., J.R.; Culhane, P.J.; Lagace, R.E.; Hicks, J.W.; Petersen,
J. Clin. Invest. 88, 1005-1013, 1991
A;Title: MUC-2 human small intestinal mucin gene structure. Repeated arrays and polym
A;Reference number: A43932; MUID:91358717
A;Accession: A43932
A;Molecule type: DNA
A;Residues: 1343-1350,'L',1352-1411,'S',1413-1448,'P',1450-1503,'T',1505-1915 <TOR>
A;Cross-references: GB:M74027; NID:g188863; PIDN:AAA59875.1; PID:g188864
A;Note: sequence inconsistent with the nucleotide translation
A;Note: sequence extracted from NCBI backbone (NCBIN:55749, NCBIP:55750)
R;Gum, J.R.; Byrd, J.C.; Hicks, J.W.; Toribara, N.W.; Lamport, D.T.A.; Kim, Y.S.
J. Biol. Chem. 264, 6480-6487, 1989
A;Title: Molecular cloning of human intestinal mucin cDNAs. Sequence analysis and evi
A;Reference number: A33532; MUID:89197956
A;Accession: B33532
A;Molecule type: mRNA
A;Residues: 1916-2193 <GU4>
A;Cross-references: GB:M22405; NID:g188873; PIDN:AAA36334.1; PID:g188874
A;Experimental source: intestine
R;Jany, B.H.; Gallup, M.W.; Yan, P.S.; Gum, J.R.; Kim, Y.S.; Basbaum, C.B.

```
Qy     837 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 882
Db     703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy     883 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 938
Db     763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy     939 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 998
Db     821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy     999 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQT 1058
Db     869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPX------------PGST 911

Qy    1059 -PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSD 1117
Db     912 APXAHGVTSAPDXRPXPGSTAPXAH----------GVTSAPDXRPXPGSTA----PXAHG 957

Qy    1118 ETNICNGKPVDGLTT-LRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGIPSPID 1172
Db     958 VTSAPDXRPXPGSTAPXAHGVTSA----------PDXRPXPGSTAPXAHGVTSAPD 1003
```

Search completed: April 26, 2002, 16:16:52
Job time: 203 sec

```
Db    912 APXAHGVTSAPDXRPXPGSTAPXAH----------GVTSAPDXRPXPGSTA----PXAHG 957

Qy   1118 ETNICNGKPVDGLTT-LRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGIPSPID 1172
Db    958 VTSAPDXRPXPGSTAPXAHGVTSA----------PDXRPXPGSTAPXAHGVTSAPD 1003


RESULT 15
US-09-083-116-2
; Sequence 2, Application US/09083116
; Patent No. 6203795
;  GENERAL INFORMATION:
;    APPLICANT:  CHAMBON, Pierre
;    APPLICANT:  KIENY, Marie-Paule
;    APPLICANT:  LATHE, Richard
;    APPLICANT:  HAREUVENI, Mara
;    TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;    TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
;    NUMBER OF SEQUENCES:  5
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
;      STREET:  P.O. Box 1404
;      CITY:  Alexandria
;      STATE:  Virginia
;      COUNTRY:  United States
;      ZIP:  22313-1404
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/09/083,116
;      FILING DATE:
;      CLASSIFICATION:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  08/479,537
;      FILING DATE:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  WO PCT/FR91/00835
;      FILING DATE:  23-OCT-1991
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/039,320
;      FILING DATE:  04-APR-1993
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/403,576
;      FILING DATE:  14-MAR-1995
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Teskin, Robin L.
;      REGISTRATION NUMBER:  35,030
;      REFERENCE/DOCKET NUMBER:  017753-025
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (703) 836-6620
;      TELEFAX:  (703) 836-2021
;  INFORMATION FOR SEQ ID NO:  2:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  2035 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  peptide
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  128..1899
;      OTHER INFORMATION:  /note= "The amino acids spanning
;      OTHER INFORMATION:  128 to 1899 constitute a repeated region wherein the repeat i
;      OTHER INFORMATION:  20 amino acids, 17 of which are fixed. The number of such
;      OTHER INFORMATION:  repeats varies from 1 to 40."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  134
;      OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;      OTHER INFORMATION:  Xaa Xaa which is the codon for Pro or Ala wherein Pro = CCT,
;      OTHER INFORMATION:  CCC, CCA, or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  144
;      OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;      OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC,
;      OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  147
;      OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC,
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  1..21
;      OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;      OTHER INFORMATION:  21 amino acid precursor sequence."
US-09-083-116-2


  Query Match           5.7%;  Score 424.5;  DB 4;  Length 2035;
  Best Local Similarity  23.7%;  Pred. No. 2.4e-18;
  Matches  255;  Conservative   69;  Mismatches  511;  Indels  241;  Gaps   53;

Qy    250 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 308
Db     16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy    309 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 351
Db     66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy    352 TTPKEPAPTTT-------KSAPTTPKEP-----------APTTTKEPAPTTPKEPAPTT 392
Db    123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy    393 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 439
Db    181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy    440 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 486
Db    241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy    487 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 535
Db    301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy    536 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 587
Db    360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy    588 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 643
Db    420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy    644 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 698
Db    472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy    699 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 746
Db    531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy    747 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 792
Db    586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy    793 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 836
Db    645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702
```

```
;    NUMBER OF SEQUENCES:  5
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
;      STREET:  P.O. Box 1404
;      CITY:  Alexandria
;      STATE:  Virginia
;      COUNTRY:  United States
;      ZIP:  22313-1404
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/08/479,537A
;      FILING DATE:  07-JUN-1995
;      CLASSIFICATION:  514
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  FR 90/13101
;      FILING DATE:  23-OCT-1990
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  WO PCT/FR91/00835
;      FILING DATE:  23-OCT-1991
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/039,320
;      FILING DATE:  04-APR-1993
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/403,576
;      FILING DATE:  14-MAR-1995
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Teskin, Robin L.
;      REGISTRATION NUMBER:  35,030
;      REFERENCE/DOCKET NUMBER:  017753-025
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (703) 836-6620
;      TELEFAX:  (703) 836-2021
;  INFORMATION FOR SEQ ID NO:  2:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  2035 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  peptide
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  128..1899
;      OTHER INFORMATION:  /note= "The amino acids spanning
;      OTHER INFORMATION:  128 to 1899 constitute a repeated region wherein the repeat i
;      OTHER INFORMATION:  20 amino acids, 17 of which are fixed. The number of such
;      OTHER INFORMATION:  repeats varies from 1 to 40."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  134
;      OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;      OTHER INFORMATION:  Xaa Xaa which is the codon for Pro or Ala wherein Pro = CCT,
;      OTHER INFORMATION:  CCC, CCA, or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  144
;      OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;      OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC, ACA
;      OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  147
;      OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  1..21
;      OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;      OTHER INFORMATION:  21 amino acid precursor sequence."
```

US-08-479-537A-2

Query Match 5.7%; Score 424.5; DB 2; Length 2035;
Best Local Similarity 23.7%; Pred. No. 2.4e-18;
Matches 255; Conservative 69; Mismatches 511; Indels 241; Gaps 53;

```
Qy   250 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 308
Db    16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy   309 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 351
Db    66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy   352 TTPKEPAPTTT-------KSAPTTPKEP-----------APTTTKEPAPTTPKEPAPTT 392
Db   123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy   393 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 439
Db   181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy   440 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 486
Db   241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy   487 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 535
Db   301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy   536 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 587
Db   360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy   588 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 643
Db   420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy   644 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 698
Db   472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy   699 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 746
Db   531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy   747 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 792
Db   586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy   793 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 836
Db   645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy   837 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 882
Db   703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy   883 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 938
Db   763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy   939 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 998
Db   821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy   999 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQT 1058
Db   869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPX----------PGST 911

Qy  1059 -PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSD 1117
```

```
;       APPLICATION NUMBER:  08/479,537
;       FILING DATE:
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  WO PCT/FR91/00835
;       FILING DATE:  23-OCT-1991
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  US 08/039,320
;       FILING DATE:  04-APR-1993
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  US 08/403,576
;       FILING DATE:  14-MAR-1995
;     ATTORNEY/AGENT INFORMATION:
;       NAME:  Teskin, Robin L.
;       REGISTRATION NUMBER:  35,030
;       REFERENCE/DOCKET NUMBER:  017753-025
;     TELECOMMUNICATION INFORMATION:
;       TELEPHONE:  (703) 836-6620
;       TELEFAX:  (703) 836-2021
;   INFORMATION FOR SEQ ID NO:  5:
;     SEQUENCE CHARACTERISTICS:
;       LENGTH:  1867 amino acids
;       TYPE:  amino acid
;       STRANDEDNESS:  single
;       TOPOLOGY:  linear
;     MOLECULE TYPE:  peptide
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  128..1727
;       OTHER INFORMATION:  /note= "The amino acids spanning
;       OTHER INFORMATION:  128 to 1727 constitute a repeated region wherein the repeat i
;       OTHER INFORMATION:  20 amino acids, 17 of which are fixed.  The number of such
;       OTHER INFORMATION:  repeats varies from 1 to 40."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  134
;       OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;       OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;       OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  144
;       OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;       OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC, ACA
;       OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  147
;       OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;       OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;       OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  1..21
;       OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;       OTHER INFORMATION:  21 amino acid precursor sequence."
US-09-083-116-5


  Query Match             5.7%;  Score 424.5;  DB 4;  Length 1867;
  Best Local Similarity   23.7%;  Pred. No. 2.1e-18;
  Matches  255;  Conservative   69;  Mismatches  511;  Indels  241;  Gaps   53;

Qy    250 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 308
          | |||          | :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db     16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy    309 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 351
          |    :||:|    |    ||              | |     ||||     |   |||
Db     66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy    352 TTPKEPAPTTT-------KSAPTTPKEP-----------APTTTKEPAPTTPKEPAPTT 392
             :||| :|         |||      |           ||     | | |     |:
Db    123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy    393 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 439
             :  | | :|         |||      |  | |        ||      |  | |     |:
Db    181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy    440 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 486
             :  | | :  |         ||      |  |||          ||      |  | |     |:
Db    241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy    487 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 535
             :  |  |   ||       |||      |  |         |||     |  |
Db    301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy    536 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 587
             ||      |  | |        ||      |  | |        ||      |  |||      |
Db    360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy    588 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 643
             :  :  |    | |        |:    ||      |  | |      |:  :  |      |  |
Db    420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy    644 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 698
             | |    :||:     |  | |      |:  :  |  |  |||      | ||      |
Db    472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy    699 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 746
             | |           :| |   ||       ||     :| |  | |      |:     ||
Db    531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy    747 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 792
             | | |      :||      |  | |           :| |    ||     | ||
Db    586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy    793 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 836
             :| |  | |       :||      |  | |           :| | :  | |      | |
Db    645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy    837 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 882
             | :       :  :||      |  |  : ::|      | :  | |
Db    703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy    883 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 938
              :|   :||      |:  | |  | :|  ||     |:     :              |||
Db    763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy    939 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 998
                |       |   | ||       :  :     | ||      |::    | |
Db    821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy    999 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQT 1058
             : ||      |  |     |       | |    :||  |:  |                | |
Db    869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPX------------PGST 911

Qy   1059 -PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSD 1117
              | :  |   |      |:      |                ||   |  |   :      |
Db    912 APXAHGVTSAPDXRPXPGSTAPXAH----------GVTSAPDXRPXPGSTA----PXAHG 957

Qy   1118 ETNICNGKPVDGLTT-LRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGIPSPID 1172
             |:  : :|  | |    :|   |          |  | |          |: |  |
Db    958 VTSAPDXRPXPGSTAPXAHGVTSA----------PDXRPXPGSTAPXAHGVTSAPD 1003


RESULT  14
US-08-479-537A-2
; Sequence 2, Application US/08479537A
; Patent No. 5861381
;  GENERAL INFORMATION:
;    APPLICANT:  CHAMBON, Pierre
;    APPLICANT:  KIENY, Marie-Paule
;    APPLICANT:  LATHE, Richard
;    APPLICANT:  HAREUVENI, Mara
;    TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;    TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
```

```
;  INFORMATION FOR SEQ ID NO:  5:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  1867 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  peptide
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  128..1727
;      OTHER INFORMATION:  /note= "The amino acids spanning
;      OTHER INFORMATION:  128 to 1727 constitute a repeated region wherein the repeat i
;      OTHER INFORMATION:  20 amino acids, 17 of which are fixed.  The number of such
;      OTHER INFORMATION:  repeats varies from 1 to 40."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  134
;      OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  144
;      OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;      OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC, ACA
;      OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  147
;      OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  1..21
;      OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;      OTHER INFORMATION:  21 amino acid precursor sequence."
US-08-479-537A-5


  Query Match             5.7%;  Score 424.5;  DB 2;  Length 1867;
  Best Local Similarity  23.7%;  Pred. No. 2.1e-18;
  Matches  255;  Conservative   69;  Mismatches  511;  Indels  241;  Gaps   53;

Qy    250 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 308
          | |||           | :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db     16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy    309 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 351
          |    :||:|    |    ||                | |     ||||     |   |||
Db     66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy    352 TTPKEPAPTTT-------KSAPTTPKEP-----------APTTTKEPAPTTPKEPAPTT 392
            :||| :|         |||      |              ||      |  | |      |:
Db    123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy    393 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 439
            :  | | :|        |||      |  | |        ||      |  | |     |:
Db    181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy    440 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 486
            :  | | :  |        ||      |  |||      ||      |  | |     |:
Db    241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy    487 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 535
            :  |  |   ||      |||      |  |          |||     |  |
Db    301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy    536 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 587
          ||      |  | |      ||      |  | |       ||      |  |||     |
Db    360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy    588 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 643
          :  :  |    | |      |:    ||      |  | |     |:  :  |     | |
Db    420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy    644 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 698
            | |    :||:    |  | |      |:  :  |  |  |||      | ||     |
Db    472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy    699 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 746
          | |           :| |   ||       ||     :| |  | |      |:    ||
Db    531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy    747 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 792
            |  | |       :||      |  | |                :| |   ||     | ||
Db    586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy    793 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 836
          :| |   | |       :||      |  | |          :| | :  | |       | |
Db    645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy    837 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 882
            | :         :  :||      |  |     : ::|     | :  | |
Db    703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy    883 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 938
           :|   :||       |:  | | |  | :|  ||      |:   :             |||
Db    763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy    939 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 998
                |       |   |  ||      :  :     |  ||      |::    | |
Db    821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy    999 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQT 1058
            : ||      |  |   |       | |    :||  |:  |                 | |
Db    869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPX------------PGST 911

Qy   1059 -PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSD 1117
           | :  |   |       |:    |                ||   |  |  :      |
Db    912 APXAHGVTSAPDXRPXPGSTAPXAH----------GVTSAPDXRPXPGSTA----PXAHG 957

Qy   1118 ETNICNGKPVDGLTT-LRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGIPSPID 1172
           |:  : :|  | |    :|   |              |  | |        |: |  |
Db    958 VTSAPDXRPXPGSTAPXAHGVTSA----------PDXRPXPGSTAPXAHGVTSAPD 1003


RESULT  13
US-09-083-116-5
; Sequence 5, Application US/09083116
; Patent No. 6203795
;  GENERAL INFORMATION:
;    APPLICANT:  CHAMBON, Pierre
;    APPLICANT:  KIENY, Marie-Paule
;    APPLICANT:  LATHE, Richard
;    APPLICANT:  HAREUVENI, Mara
;    TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;    TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
;    NUMBER OF SEQUENCES:  5
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
;      STREET:  P.O. Box 1404
;      CITY:  Alexandria
;      STATE:  Virginia
;      COUNTRY:  United States
;      ZIP:  22313-1404
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/09/083,116
;      FILING DATE:
;      CLASSIFICATION:
;    PRIOR APPLICATION DATA:
```

; FILING DATE:
; CLASSIFICATION:
; ATTORNEY/AGENT INFORMATION:
; NAME: Campbell, Cathryn A.
; REGISTRATION NUMBER: 31,815
; REFERENCE/DOCKET NUMBER: P-LJ 2626
; TELECOMMUNICATION INFORMATION:
; TELEPHONE: (619) 535-9001
; TELEFAX: (619) 535-8949
; INFORMATION FOR SEQ ID NO: 23:
; SEQUENCE CHARACTERISTICS:
; LENGTH: 1185 amino acids
; TYPE: amino acid
; TOPOLOGY: linear
; MOLECULE TYPE: protein
US-09-041-886-23

Query Match 5.7%; Score 425.5; DB 4; Length 1185;
Best Local Similarity 21.8%; Pred. No. 1.1e-18;
Matches 240; Conservative 119; Mismatches 437; Indels 303; Gaps 52;

```
Qy    102 SQTIKSTTKRSPKPPNKKKTKKVIESEEITEEHSVSENQESSSSSSSSSSSTIWKIKSS 161
          | :::|  |:    |           ||:      |    |:|||              | : |
Db      9 SMSMRSGRKKEAPGPR----------EELRSRGRASPGGVSTSSSDG--------KAEKS 50

Qy    162 KNSAANRELQKKLKVKDNKKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNK 221
          : :|              || | ::  |||   |:: |                :  :  ::
Db     51 RQTA-------------KKARVEEASTPK---VNKQGRS----------EEISESESE 82

Qy    222 VSTSPKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTT 281
          : :|| |  :   |||  | : |:    ||  | : : : ::    |: ||
Db     83 ETNAPKKTKTEQELPRPQSPSDLDSLDGRSL--NDDGSSDPRDIDQDNRSTS-------P 133

Qy    282 SAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTP--KEPASTTP 339
          |     |:|  |              :  ::|||   |  | |  |   :|  :||
Db    134 SIYSPGSVENDS--------------DSSSGLSQGPA--RPYHPPPLFPPSPQPPDSTP 176

Qy    340 KEPTPTTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPT 399
          ::|  :          || |: | :    | || ||:    ||  |  | |      |  |
Db    177 RQPEASF---------EPHPSVTPTGYHAPMEP-PTSRMFQAP--PGAPPPHPQLYPGGT 224

Qy    400 --TTKSAPTTPK---------------EPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPK 441
              |  ||                 | ||||   : :      || |    ||||
Db    225 GGVLSGPPMGPKGGGAASSVGGPNGGKQHPPPTTPISVSSSGASGAPPTKP----PTTPV 280

Qy    442 EPAPTTKEPAP-----TTPKEPAPTA--PKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTT 494
              | |       ||  | | |  |   | :|    |     :| |      |||
Db    281 GGGNLPSAPPPANFPHVTPNLPPPPALRPLNNASASP----PGLGAQPLPGHL--PSPYA 334

Qy    495 PKE------PAPTTTKSAPTTTKEP--APTTTKSAPTTP-KEPSPTTTKEPAPTTPKEPA 545
            :      | |   :  ||     |   |  : |||  | : |  :::   |  :   :
Db    335 MGQGMGGLPPGP---EKGPTLAPSPHSLPPASSSAPAPPMRFPYSSSSSSSAAASSSSSS 391

Qy    546 PTTPKKPAPTTPKEPAPTTPKE-PAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLT--- 601
            ::   | | :  :  |: |   | ||:      :   :|    || | |: | :
Db    392 SSSSASPFPAS--QALPSYPHSFPPPTSL-----SVSNQP----PKYTQPSLPSQAVWSQ 440

Qy    602 --PTTPE--KLAPTTPEKPAPTTPEELAPTTPEEPTPT---------------------- 635
              |  |   :|   :    | |  |    | :|    |  |
Db    441 GPPPPPPYGRLLANSNAHPGPFPPSTGAQSTAHPPVSTHHHHHQQQQQQQQQQQQQQHHG 500

Qy    636 --TPEEPAP-----------TTPKAAAPN----TPKEPAPTTPKEP------APTTPKE 671
              |  |              | | :|:      |  | |  |       :   |
Db    501 NSGPPPPGAFPHPLEGGSSHHAHPYAMSPSLGSLRPYPPGPAHLPPPHSQVSYSQAGPNG 560

Qy    672 PAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAP---TTTKEPTSTTSDKPAPT 728
            |  ::   :: :| :|: | :   |:|:   : ||    |   ||    | :|    :
Db    561 PPVSSSSNSSSSTSQGSYPCS--HPSPSQGPQGAPYPFPPVPTVTTSSATLSTVIATVAS 618

Qy    729 TPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTS 788
            :| |    :|  | |   : |:|  |   |   |  :|     |: :   |   :|
Db    619 SPAGYKTASPPGPPPYGKRAPSPGAYKTATPPGYKPGSP-----PSFRTGTPPGYRG--- 670
```

```
Qy    789 TTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKS 848
            ||     | | |  :||    | |  |  |:   |  |||  :  | |  :    | | :
Db    671 -TSPPAGPGTFKPGSPTV--GPGPLPPAGPS-GLPSLPPPPAAPASGPPLS--ATQIKQE 724

Qy    849 PDE--STPELSAEPTPKALENS--PKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTP 904
            | |  |||    | | |   |  ||   ||:  : :|   :      :    ||   |
Db    725 PAEEYETPE---SPVPPARSPSPPPKVVDVPSHASQSARFNKHLDRGFNSCARSDLYFVP 781

Qy    905 ETTTAAPKMTKETATTTEKT-TESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTT 963
                  |: |: |   ||   |::  |
Db    782 ---LEGSKLAKKRADLVEKVRREAEQRAR------------------------------ 807

Qy    964 TKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPK 1023
                        ||  : ::|    :    | ::  :: |    : |    | |
Db    808 -------------EEKERERERERERERERERERELERSVKLAQEGRAP-----VECPS 848

Qy   1024 TTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHM 1083
            | | :     |   | | :|      |  |  |::   :     || |
Db    849 LGPVPHR-----PPFEPGSAVA------TVPPYLGPDTPALRT--LSEYA---------- 885

Qy   1084 LLRPHVFMPEVTPDMDYLP 1102
            ||||  |       |:|
Db    886 --RPHVMSPGNRNHPFYVP 902
```

RESULT 12
US-08-479-537A-5
; Sequence 5, Application US/08479537A
; Patent No. 5861381
; GENERAL INFORMATION:
; APPLICANT: CHAMBON, Pierre
; APPLICANT: KIENY, Marie-Paule
; APPLICANT: LATHE, Richard
; APPLICANT: HAREUVENI, Mara
; TITLE OF INVENTION: PHARMACEUTICAL COMPOSITION FOR THE
; TITLE OF INVENTION: TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
; NUMBER OF SEQUENCES: 5
; CORRESPONDENCE ADDRESS:
; ADDRESSEE: BURNS, DOANE, SWECKER & MATHIS, L.L.P.
; STREET: P.O. Box 1404
; CITY: Alexandria
; STATE: Virginia
; COUNTRY: United States
; ZIP: 22313-1404
; COMPUTER READABLE FORM:
; MEDIUM TYPE: Floppy disk
; COMPUTER: IBM PC compatible
; OPERATING SYSTEM: PC-DOS/MS-DOS
; SOFTWARE: PatentIn Release #1.0, Version #1.30
; CURRENT APPLICATION DATA:
; APPLICATION NUMBER: US/08/479,537A
; FILING DATE: 07-JUN-1995
; CLASSIFICATION: 514
; PRIOR APPLICATION DATA:
; APPLICATION NUMBER: FR 90/13101
; FILING DATE: 23-OCT-1990
; PRIOR APPLICATION DATA:
; APPLICATION NUMBER: WO PCT/FR91/00835
; FILING DATE: 23-OCT-1991
; PRIOR APPLICATION DATA:
; APPLICATION NUMBER: US 08/039,320
; FILING DATE: 04-APR-1993
; PRIOR APPLICATION DATA:
; APPLICATION NUMBER: US 08/403,576
; FILING DATE: 14-MAR-1995
; ATTORNEY/AGENT INFORMATION:
; NAME: Teskin, Robin L.
; REGISTRATION NUMBER: 35,030
; REFERENCE/DOCKET NUMBER: 017753-025
; TELECOMMUNICATION INFORMATION:
; TELEPHONE: (703) 836-6620
; TELEFAX: (703) 836-2021

```
Qy    454 TPKEPAPTAPKKPA--PTTPKEPA--PTTPKEPAPTTTKEPSPTTP-----KEPAPTTTK 504
Db    197 YP----PTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPSLIKAKPSYPPTYK 252

Qy    505 SAPTTTKEPA-------PTTTKSAPTTPKEPSPTTTKEPA--PTTPKEPA----PTTPKK 551
Db    253 AKPTYKAKPTYKAKPTYPSTYKAKPTYP----PTYKAKPSYPPTYKAKPTYKAKPTYKAK 308

Qy    552 PA--PTTPKEPA--PTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEK 607
Db    309 PSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYP----PTYKAKPSYPPTYKVKPT--YK 362

Qy    608 LAPTTPE--KPAPTTPEELAPTTPEEPT--PTTPEEPA-PTTPKAAAPNTPKEPAPTTPK 662
Db    363 AKPTYPSTYKAKPSYP----PTYKAKPSYPPTYKAKPSYPPTYK-AKPSYP----PTYKA 413

Qy    663 EPA-PTT----PKEPAPTTPKETAPTT--PKGTAPTTLKEPAPTTPKKPAPKELAPTTTK 715
Db    414 KPTYPSTYKAKPSYPPTYKPKISYPPTYKAKPSYPSTYKAKSSYPPTYKAKPSYPPTYKA 473

Qy    716 EPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPA--PTTPKKP 773
Db    474 KPTYPSTYKAKPTYK--AKPSYP----PTYKAKPS------YPPTYKAKPSYPPTYKAKP 521

Qy    774 APKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPA--PTTPKKPA-PTT----PETP 826
Db    522 TYK-AKPTYKAKPSYPPTYKAKPSYP----PTYKAKPSYPPTYKAKPTYPSTYKAKPSYP 576

Qy    827 PPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPE 886
Db    577 PTYKAKPSYPPT-------YKAKPSYPPTYKAKPTYKA------KPTYPST---YKAKPS 620

Qy    887 MTTTAKDKTTERDLRTTPETTTAAP 911
Db    621 YPPTHKAKPTYKAKPSYPPTHKAKP 645
```

RESULT 10
US-09-219-849-5
; Sequence 5, Application US/09219849
; Patent No. 6150081
; GENERAL INFORMATION:
; APPLICANT: VAN HEERDE, GEORGE V.
; APPLICANT: VAN RIJN, ALEXIS C.
; APPLICANT: BOUWSTRA, JAN B.
; APPLICANT: DE WOLF, FREDERIK A.
; APPLICANT: MOOBROEK, ANDREAS
; APPLICANT: WERTEN, MARC W.T.
; APPLICANT: WIND, RICHELE D.
; APPLICANT: VAN DEN BOSCH, TANJA J.
; TITLE OF INVENTION: SILVER HALIDE EMULSIONS WITH RECOMBINANT COLLAGEN
; TITLE OF INVENTION: SUITABLE FOR PHOTOGRAPHIC APPLICATION AND ALSO THE
; TITLE OF INVENTION: PREPARATION THEREOF
; FILE REFERENCE: 2728-2
; CURRENT APPLICATION NUMBER: US/09/219,849
; CURRENT FILING DATE: 1998-12-23
; NUMBER OF SEQ ID NOS: 50
; SOFTWARE: PatentIn Ver. 2.1
; SEQ ID NO 5
; LENGTH: 960
; TYPE: PRT
; ORGANISM: Artificial Sequence
; FEATURE:
; OTHER INFORMATION: Description of Artificial Sequence: Illustrative
; OTHER INFORMATION: amino acid sequence
US-09-219-849-5

Query Match 6.1%; Score 452; DB 4; Length 960;
Best Local Similarity 25.6%; Pred. No. 1.9e-20;
Matches 173; Conservative 60; Mismatches 309; Indels 134; Gaps 41;

```
Qy    307 PTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAP--TTTKS 364
Db     72 PGPAGPPGSRDPG--PPGAPGPAGP--PGSRDPGPPGAPG-PAGPPGSRDPGPPGAPGPA 126

Qy    365 APTTPKEPAPTTTKEPA-PTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTP---KKPA 420
Db    127 GPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPG 186

Qy    421 PTTPKEPAPTTP---KEPTPTTPKEPA--PTTKEPAPTTPKEPA--------------- 459
Db    187 P--PGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAHGPAGPKGAHGPAGPKGAHG 244

Qy    460 ---------PTAPK-KPAPTTP---KEPAPTTPKEPA-PTTTKEPSPTTPKEPAPTTTKS 505
Db    245 PAGPKGAHGPAGPKGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP--PGAPGP----A 298

Qy    506 APTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTP---KKPAPTTPKEPAP 562
Db    299 GPPGSRDPGP------PGAPGPAGPPGSRDPGP--PGAPGPAGPPGSRDPGP--PGAPGP 348

Qy    563 TTP---KEPAPTTTKKPA-------PTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTT 612
Db    349 AGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP----PGAPGPAGPPG 404

Qy    613 PEKPAPT-TPEELAPTTPEEPTPTTPEEPAPTTPKAAA-PNTPKEPAPTTP---KEPAPT 667
Db    405 SRDPGPPGAPGPAGPPGSRDPGP--PGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPP 462

Qy    668 TPKEPA-------PTTPKET-APTTPKGT--------APTTLKEPAPTTPKKP-APKELA 710
Db    463 GAHGPAGPKGAHGPAGPKGAHGPAGPKGAHGPAGPKGAPGPAGPPGSRDPGPPGAPGPAG 522

Qy    711 PTTTKEPTSTTSDKPAPTTPKGTA-PTTPKEPAPTTP---KEPAPTTPKGTA-PTTLKEP 765
Db    523 PPGSRDPGPPGA--PGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDP 580

Qy    766 APTTPKKPAP------KELAPTTTKGPTST--TSDKPAPTTPKETAPTTPKEPAPTTPKK 817
Db    581 GP--PGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP--PGA 636

Qy    818 PAPTTP----ETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALEN-----S 868
Db    637 PGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP-PGAPGPAGPPGSRDPGPPGAPGPAGPPG 695

Qy    869 PKEPGVPTTKTPAATK 884
Db    696 SRDPGPPGAHGPAGPK 711
```

RESULT 11
US-09-041-886-23
; Sequence 23, Application US/09041886
; Patent No. 6235872
; GENERAL INFORMATION:
; APPLICANT: Bredesen, Dale E.
; APPLICANT: Rabizadeh, Sharroz
; TITLE OF INVENTION: Proapoptotic Peptides, Dependence
; TITLE OF INVENTION: Polypeptides and Methods of Use
; NUMBER OF SEQUENCES: 72
; CORRESPONDENCE ADDRESS:
; ADDRESSEE: Campbell & Flores LLP
; STREET: 4370 La Jolla Village Drive, Suite 700
; CITY: San Diego
; STATE: California
; COUNTRY: United States
; ZIP: 92122
; COMPUTER READABLE FORM:
; MEDIUM TYPE: Floppy disk
; COMPUTER: IBM PC compatible
; OPERATING SYSTEM: PC-DOS/MS-DOS
; SOFTWARE: PatentIn Release #1.0, Version #1.25
; CURRENT APPLICATION DATA:
; APPLICATION NUMBER: US/09/041,886

```
Qy    686 KGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTT 745
                  |: |::| |:|  |:  |||    |:   |  |:   |  |  |  :
Db    565 -----------PSNPEEPNPEE--PSNPKEP----SNPEEPINPEELNPKEPSNPEESN 606

Qy    746 PKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPT 805
          ||||              |  : ||:|    : |   : |      :   |       |
Db    607 PKEPI------------NPEESNPKEP----INPEDNENPLIIQDEPIEPRNDSNVIPI 649

Qy    806 TPKEP------APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAE 859
          |  |          |: |: :  | | |  :  ::  | |     |   |:|  |
Db    650 LPIIPQKGNNIPSNLPENPSDSEVEYPRPNDNGENSNNTMKSKKNI---PNEPIPSPGDN 706

Qy    860 PTPKALENSPK-------------------EPGVPTTK 878
          |     |  ||                    || :|  :
Db    707 PYKGHEERIPKPHRSNDDYVYDNNVNKNNKDEPEIPNNE 745
```

RESULT 8
PCT-US92-00018-2
; Sequence 2, Application PC/TUS9200018
; GENERAL INFORMATION:
; APPLICANT: Hoffman, Stephen L.
; APPLICANT: Charoenvit, Yupin
; APPLICANT: Hedstrom, Richard
; APPLICANT: Khusmith, Srisin
; APPLICANT: Rogers IV, William O.
; TITLE OF INVENTION: Protective malaria sporozoite surface protein
; TITLE OF INVENTION: immunogen and gene encoding
; NUMBER OF SEQUENCES: 2
; CORRESPONDENCE ADDRESS:
; ADDRESSEE: A. David Spevack
; STREET: NMRDC Building 1 T-12 National Naval
; STREET: Medical Center
; CITY: Bethesda
; STATE: MD
; COUNTRY: USA
; ZIP: 20814-5044
; COMPUTER READABLE FORM:
; MEDIUM TYPE: Floppy disk
; COMPUTER: IBM PC compatible
; OPERATING SYSTEM: PC-DOS/MS-DOS
; SOFTWARE: PatentIn Release #1.24
; CURRENT APPLICATION DATA:
; APPLICATION NUMBER: PCT/US92/00018
; FILING DATE: 19920103
; CLASSIFICATION: 424
; ATTORNEY/AGENT INFORMATION:
; NAME: Spevack, Avram D.
; TELECOMMUNICATION INFORMATION:
; TELEPHONE: (301) 295-6759
; TELEFAX: (301) 295-4033
; INFORMATION FOR SEQ ID NO: 2:
; SEQUENCE CHARACTERISTICS:
; LENGTH: 826 amino acids
; TYPE: AMINO ACID
; TOPOLOGY: linear
; MOLECULE TYPE: protein
PCT-US92-00018-2

Query Match 6.6%; Score 488.5; DB 5; Length 826;
Best Local Similarity 26.4%; Pred. No. 8.9e-23;
Matches 153; Conservative 54; Mismatches 223; Indels 149; Gaps 34;

```
Qy    343 TPTTIKSAPTTPKEPA-PTTTKSAPTTPKEPA-PTTTKEP----APTTPKEP-APTTTKE 395
          ||  ::  |  |  |  |       |: |:|| |    :|     |  |  |  |
Db    273 TPCKVRDCPQIPIPPVIPNKIPEKPSNPEEPVNPNDPNDPNNPNNPNNPNNPNNPNN 332

Qy    396 PAPTTTKSAPTTPKEP-APTTPKKP-APTTPKEP-APTTPKEP-TPTTPKEPAPTTKEPA 451
          |      : |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |:|:      |
Db    333 PNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNDPSNPNNHPK 392

Qy    452 PTTPKEPAPTAPK--KPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTT 509
```

```
Db    393 RRNPKRRNPNKPKPNKPNPNKPNPNEPSNPNKPNPN-----EPSNPNKPNP----NEPSN 443

Qy    510 TKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA-PTTPKKP-APTTPKEPAPTTPKE 567
            :| |    : |: | :|:|       |  | ||: |  |  |  || : |  |  |: | |
Db    444 PNKPNP----NEPSNPNKPNPN-----EPLNPNEPSNPNEPSNPNAPSNPNE--PSNPNE 492

Qy    568 PAPTTTKKPAPTAPKEP-APTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAP 626
          |:          |  |  |  |: | |  |: |||                    |: | |  |
Db    493 PS-------NPNEPSNPNEPSNPNE--PSNPKK-----------------PSNPNE--P 523

Qy    627 TTPEEPTPTTPEEPAPTTPKAAAPNTPKEP-APTTPKEPAPTTPKEPAPTTPKETAPTTP 685
          : | |  |  | ||       : || |  |  |: |:|  |: |||  |: | |
Db    524 SNPNE--PLNPNEP-------SNPNEPSNPNEPSNPEE--PSNPKE--PSNPNE------ 564

Qy    686 KGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTT 745
                      |: |::| |:|  |:  |||    |:   |  |:   |  |  |  :
Db    565 -----------PSNPEEPNPEE--PSNPKEP----SNPEEPINPEELNPKEPSNPEESN 606

Qy    746 PKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPT 805
          ||||              |  : ||:|    : |   : |      :   |       |
Db    607 PKEPI------------NPEESNPKEP----INPEDNENPLIIQDEPIEPRNDSNVIPI 649

Qy    806 TPKEP------APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAE 859
          |  |          |: |: :  | | |  :  ::  | |     |   |:|  |
Db    650 LPIIPQKGNNIPSNLPENPSDSEVEYPRPNDNGENSNNTMKSKKNI---PNEPIPSPGDN 706

Qy    860 PTPKALENSPK-------------------EPGVPTTK 878
          |     |  ||                    || :|  :
Db    707 PYKGHEERIPKPHRSNDDYVYDNNVNKNNKDEPEIPNNE 745
```

RESULT 9
5202236-13
;Patent No. 5202236
; APPLICANT: MAUGH, KATHY J.;ANDERSON, DAVID M.;STRAUSBERG,
;SUSAN L.;MCCANDLISS, RUSS;WEI, TENA;FILPULA, DAVID
; TITLE OF INVENTION: METHOD OF PRODUCING BIOADHESIVE
;PROTEIN
; NUMBER OF SEQUENCES: 39
; CURRENT APPLICATION DATA:
; APPLICATION NUMBER: US/07/528,762
; FILING DATE: 25-MAY-1990
; APPLICATION NUMBER: 82,456
; FILING DATE: 07-AUG-1987
; APPLICATION NUMBER: 933,945
; FILING DATE: 24-NOV-1986
; APPLICATION NUMBER: 650,128
; FILING DATE: 13-SEP-1984
;SEQ ID NO:13:
; LENGTH: 652
5202236-13

Query Match 6.4%; Score 476.5; DB 6; Length 652;
Best Local Similarity 29.7%; Pred. No. 3.7e-22;
Matches 221; Conservative 71; Mismatches 284; Indels 169; Gaps 45;

```
Qy    226 PKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKE 285
          ||:|'     |:|| ||                | ::| |                | | |
Db     11 PKMTYPPTYKPKPSYPP---------------TYKSKPTY----------KPKIT---- 40

Qy    286 TQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKE----PASTTPKE 341
                  || |  |||:         |    ||:  | | |      | :  ||
Db     41 ----------YPPTYK--AKPS--------YPPTYKPKKTYPPTYKPKLTYPPTYKPKP 79

Qy    342 PTPTTIKSAPT-TPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPA--P 398
            | | || ||  ||   | | |: |: |    |  | |    ||   | | | |  | |
Db     80 SYPPTYKSKPTYKPKITYPPTYKAKPSYPPTYKPKKTYPPT-YKPKLTYPPTYKPKASYP 138

Qy    399 TTTKSAPTTPKEPAPTTPKKPAPT-TPKEPAPTT----PKEPTPTTPKEPAPTTKEPAPT 453
          | |  |: |  |:  | |   ||  ||  | |    | |    | :  | | :|  |
Db    139 PTYKPKPSYP--PSYKTKKTYPPTYKPKLTYPPTYKPKPSYPPSYKPNKTYPPTYKPKLT 196
```

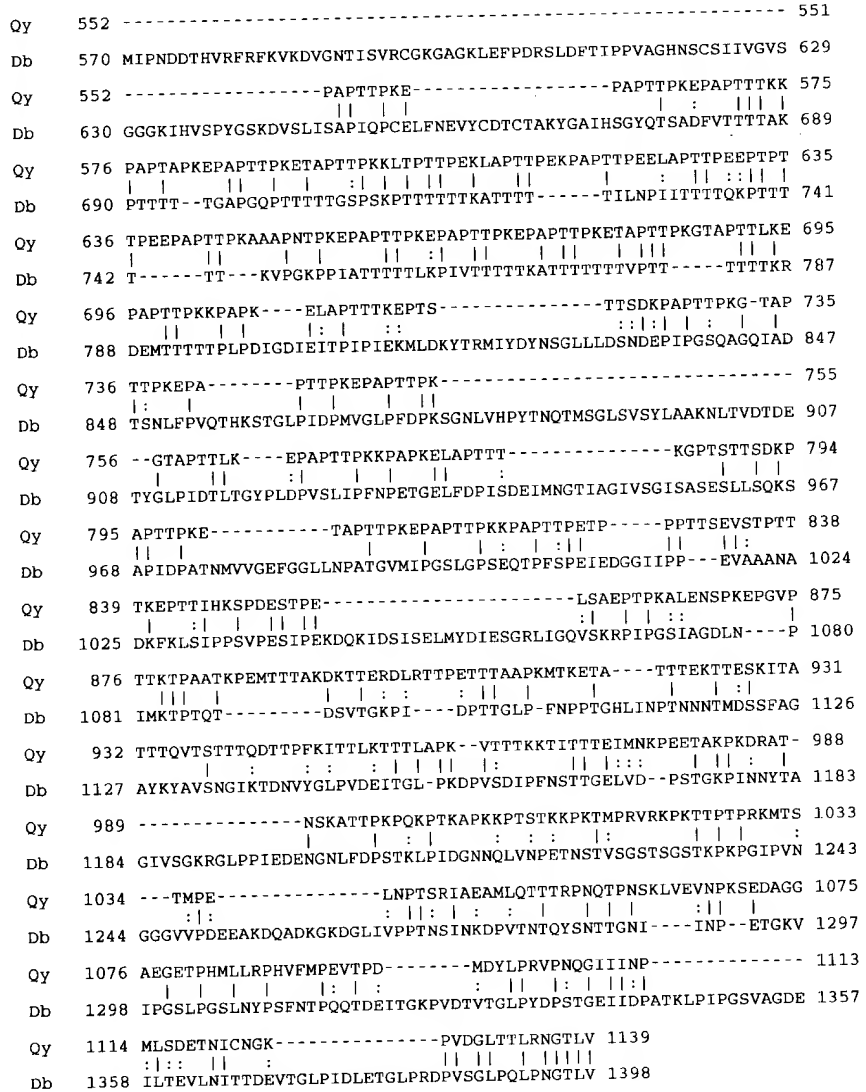
```
Qy    376 TTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEP 435
Db    272 TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTKKPTTTTTTTTTTTTTTTTTTTTTTT 331

Qy    436 TPTTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTP 495
Db    332 TTTTTKKPTTTTTTTTTT-------------TTTTTTTTTTTTTTTTTTTTTTTTTTT 377

Qy    496 KEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKE---------------------------- 527
Db    378 TKPTTTTTTTTTTTTKKPTTTTTATTTTTTSETESVIKPDEWCWLEKNGECEAKGATYVG 437

Qy    528 ----------------PSPTT---------------------------------TKE 535
Db    438 VIGKDGRIENGMAFTMIPNDDTHVRFRFKVKDVGNTISVRCRKGAGKLEFPDRSLDFTIP 497

Qy    536 P--------------------APTTPKEPAPTTPKKPAPTTPKE--------------- 559
Db    498 PVAGHNSCSIIVGVSGDGKIHVSPYGSKDVSLIS----APIQPSELFNEVYCDTCTAKYG 553

Qy    560 ---PAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKP 616
Db    554 AIHSGYQTSADFVTTTTAKPTTTT--TGAPGQPTTTTTGSPSKPTTTTTTKATTTT---- 607

Qy    617 APTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTT 676
Db    608 --TTLNPIITTTTQKPTTTT------TT---KVPGKPPIATTTTTLKPIVTTTTTKATTT 656

Qy    677 PKETAPTTPKGTAPTTLKEPAPTTPKKPAPK----ELAPTTTKEPTS------------ 719
Db    657 TTTTVPTT-----TTTTKRDEMTTTTTPLPDIGDIEITPIPIEKMLDKYTRMIYDYNSGL 711

Qy    720 --TTSDKPAPTTPKG-TAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKE---PAPTTPKKP 773
Db    712 LLDSNDEPIPGSQAGQIADTSNLFPVQTHKSTGLPIDPMVGLPFDPKSGNLVHPYTNQTM 771

Qy    774 APKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEV 833
Db    772 SGLSVSYLAAKNLTVDTDETYG--LPIDTLTGYPLDPVSLIPFN--PETGELFDPISDEI 827

Qy    834 STPT--------TTKEPTTIHKS----------------------------PDEST- 853
Db    828 MNGTIAGIVSGISASESLLSQKSALIDPATNMVVGEFGGLLNPATGVMIPGFLGPSEQTQ 887

Qy    854 ------------PELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTER--- 898
Db    888 FSPEIEDGGIIPPEVAAANADKFKLSIP--PSVP------ESIPE-----KDQKIDSISE 934

Qy    899 --------------DLRTTPETTTAAPKMTKETATTTEKTTESKITATT-TQVTSTTTQD 943
Db    935 LMYDIESGRLIGQVSKRPIPGSIAGDLNPIMKTPTQTDSVTGKPIDPTTGLPFNPPTGHL 994

Qy    944 TTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKP--EETAKPKD-------RATNSKATT 994
Db    995 INPTNNNTMDSSFAGAYKYAVSNGIKTDNVYGLPVGEITGLPKDPGSDIPFNSTTGELVD 1054

Qy    995 PKPQKPTKAPKKPTSTKKPKTMP--------------------RVRKPKTTPTPRKMTS 1033
Db   1055 PSTGKPINNSTAGIVSGKPGLPPIEDENGNLFDPSTNLPIDGNNQLVNPETNSTVSGSTS 1114

Qy   1034 --TMPE---------------------------LNPTSRIAEAMLQTTTRPNQTPNSKL 1063
Db   1115 GTTKPKPGIPVNGGGVVPDEEAKDQADKGKDGLIVPPTNSINKDPVTNTQYSNTTGNI-- 1172

Qy   1064 VEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPD--------MDYLPRVPNQGIIINP-- 1113
Db   1173 --INP--ETGKVIPGSLPGSLNYPSFNTPQQTDEITGKPVDTVTGLPYDPSTGEIIDPAT 1228

Qy   1114 -----------MLSDETNICNGK--------------PVDGLTTLRNGTLV 1139
Db   1229 KLPIPGSVAGDEILTEVLNITTDEVTGLPIDLETGLPRDPVSGLPQLPNGTLV 1281
```

```
RESULT   7
US-07-638-431-2
; Sequence 2, Application US/07638431
; Patent No. 5198535
;  GENERAL INFORMATION:
;    APPLICANT:  Hoffman, Stephen L.
;    APPLICANT:  Charoenvit, Yupin
;    APPLICANT:  Hedstrom, Richard
;    APPLICANT:  Khusmith, Srisin
;    APPLICANT:  Rogers IV, William O.
;    TITLE OF INVENTION:  Protective malaria sporozoite surface protein
;    TITLE OF INVENTION:  immunogen and gene
;    NUMBER OF SEQUENCES:  2
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  A. David Spevack
;      STREET:  NMRDC Building 1 T-12 National Naval
;      STREET:  Medical Center
;      CITY:  Bethesda
;      STATE:  MD
;      COUNTRY:  USA
;      ZIP:  20814-5044
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.24
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/07/638,431
;      FILING DATE:  19910110
;      CLASSIFICATION:  424
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Spevack, Avrom D.
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (301) 295-6759
;      TELEFAX:  (301) 295-4033
;  INFORMATION FOR SEQ ID NO:  2:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  826 amino acids
;      TYPE:  AMINO ACID
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
US-07-638-431-2


  Query Match             6.6%;  Score 488.5;  DB 1;  Length 826;
  Best Local Similarity  26.4%;  Pred. No. 8.9e-23;
  Matches  153;  Conservative   54;  Mismatches  223;  Indels  149;  Gaps   34;

Qy    343 TPTTIKSAPTTPKEPA-PTTTKSAPTTPKEPA-PTTTKEP----APTTPKEP-APTTTKE 395
Db    273 TPCKVRDCPQIPIPPVIPNKIPEKPSNPEEPVNPNDPNDPNNPNNPNNPNNPNNPNNPNN 332

Qy    396 PAPTTTKSAPTTPKEP-APTTPKKP-APTTPKEP-APTTPKEP-TPTTPKEPAPTTKEPA 451
Db    333 PNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNDPSNPNNHPK 392

Qy    452 PTTPKEPAPTAPK--KPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTT 509
Db    393 RRNPKRRNPNKPKPNKPNPNKPNPNEPSNPNKPNPN-----EPSNPNKPNP----NEPSN 443

Qy    510 TKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA-PTTPKKP-APTTPKEPAPTTPKE 567
Db    444 PNKPNP----NEPSNPNKPNPN-----EPLNPNEPSNPNEPSNPNAPSNPNE--PSNPNE 492

Qy    568 PAPTTTKKPAPTAPKEP-APTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAP 626
Db    493 PS-------NPNEPSNPNEPSNPNE--PSNPKK-----------------PSNPNE--P 523

Qy    627 TTPEEPTPTTPEEPAPTTPKAAAPNTPKEP-APTTPKEPAPTTPKEPAPTTPKETAPTTP 685
Db    524 SNPNE--PLNPNEP-------SNPNEPSNPNEPSNPEE--PSNPKE--PSNPNE------ 564
```

```
;      CLASSIFICATION:  800
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Michaels, Christopher A
;      REGISTRATION NUMBER:  34,390
;      REFERENCE/DOCKET NUMBER:  BTI-39
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (607) 256-2000
;      TELEFAX:  (607) 256-3628
;  INFORMATION FOR SEQ ID NO:  3:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  786 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
;    HYPOTHETICAL:  NO
;    ANTI-SENSE:  NO
;    ORIGINAL SOURCE:
;      ORGANISM:  Trichoplusia ni
;      TISSUE TYPE:  peritrophic membrane
US-09-103-429A-3


  Query Match             6.7%;  Score 498.5;  DB 4;  Length 786;
  Best Local Similarity   29.1%;  Pred. No. 2e-23;
  Matches  200;  Conservative   25;  Mismatches  207;  Indels  255;  Gaps   32;

Qy    307 PTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAP 366
          | | || ||: || |      ||| : | :||          ||||  :   |||  ||
Db     93 PGPPAE-TTQAPATT----QAPTTTQAPTTTT----------QAPTTTTQ--ATTTTQAP 135

Qy    367 TTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKE 426
          || :  |||||:  | ||    ||||| : |||||  ||||  : |||| :  |||| :
Db    136 TTTQ--APTTTQ--ATTT---QAPTTTTQ-APTTTTQAPTTTTQ-APTTTQ--APTTTQ- 183

Qy    427 PAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTT-PKEPA-- 483
           |||| : ||  |     |    || |||    | |   || |||  ||||: |  |
Db    184 -APTTTQAPTTITQAATTPAATTPAATTPAATTPAA-TTPAATTPGVPAPTSAPVWPPIC 241

Qy    484 ----------------------------------------------PTTTKEPSPTTPK- 496
                                                         |   :  ||   :
Db    242 ELLPNGCPADFDIHLLIPHDKYCNLFYQCSNGYTFEQRCPEGLYFNPYVQRCDSPANVEC 301

Qy    497 ----EPAPTTT---------------------------------------------- 503
               |||  |
Db    302 DGEISPAPPVTEGNEDEDIDIGDLLDNGCPANFEIDWLLPHGNRCDKYYQCVHGNLVERR 361

Qy    504 --------------------------KSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEP 536
                                     :|     | |  |    || | |  |      |
Db    362 CGAGTHFSFELQQCDHIELVGCTLPGGESEEVDVDEDA-CTGWYCPTEPIEWEPLPNGCP 420

Qy    537 APTT-----PKEP-----------------------APTTPKKPAPTT----------- 556
          |  :     | |                       :| |    :| |
Db    421 ADFSIDHLLPHESDCGQYLQCVHGQTIARPCPGNLHFSPATQSCESPVTAGCQVFECDSD 480

Qy    557 ---PKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTP 613
                |||     |||     ||||   |:   |  | | |   : |||     ||
Db    481 NQCTSTAAPTAAPTAAPTAAPTAAPTA--APSTVVPPATPPATAAPVPPTT---AIPT-- 533

Qy    614 EKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNT--PKEPAPTTPKEPAPTTPKE 671
           ||||     ||||    :|||   | ||    ||| |  |: |   |    |||
Db    534 --PAPTAAPTAAPTTAAPESPTTVTVP-PTAAPTAAPTTAVPEIPITVT---SAPTAAPT 587

Qy    672 PAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPK 731
            |||     ||||     |||:  |    |   ||   || ||              | |
Db    588 AAPTAAPTAAPTTAVPEIPTTVTSPPTAAPTTAAP---APNTT-----------VTVPP 632

Qy    732 GTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTS 791
           ||||   ||| |     ||     || |:                 || ||  | :|||
Db    633 TAAPTT-AAPAPNTTVTVPPTAAPTAAPPTVAH-------------APNTTAAPVTTTS 677

Qy    792 DKPAPTTPKETAPTTPKEPAPTTPKKP 818
             || |  |     |  | |  |  |
```

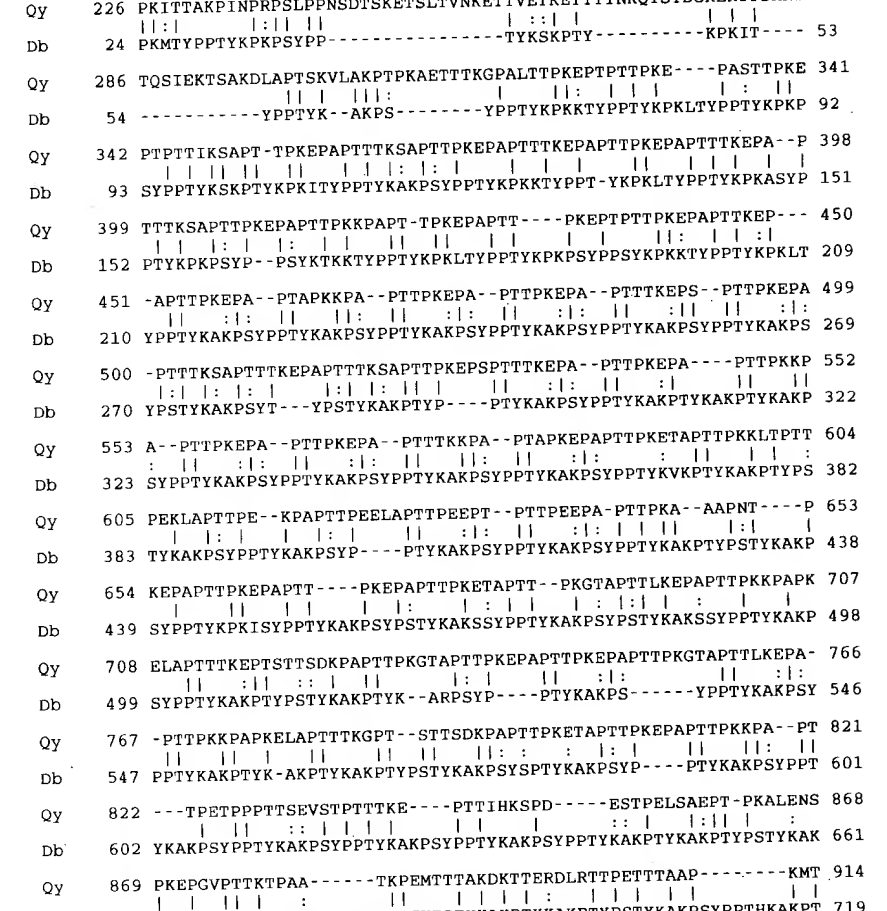
```
Db    678 ---APATTPEDDDIDP--PLPNDPINP 699


RESULT   6
US-08-928-361B-6
; Sequence 6, Application US/08928361B
; Patent No. 6071518
;  GENERAL INFORMATION:
;    APPLICANT:  Petersen, Carolyn
;    TITLE OF INVENTION:  PEPTIDES, POLYPEPTIDES, GLYCOPROTEINS,
;    TITLE OF INVENTION:  THEIR FUNCTIONAL MUTANTS, VARIANTS, ANALOGS AND FRAGMENTS
;    TITLE OF INVENTION:  FOR TREATMENT AND DETECTION/DIAGNOSIS OF CRYPTOSPORIDIUM
;    TITLE OF INVENTION:  SPECIES INFECTIONS
;    NUMBER OF SEQUENCES:  30
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  PETERS, VERNY, JONES & BIKSA
;      STREET:  385 Sherman Avenue, Suite 6
;      CITY:  Palo Alto
;      STATE:  CA
;      COUNTRY:  USA
;      ZIP:  94306-1840
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/08/928,361B
;      FILING DATE:  12-SEP-1997
;      CLASSIFICATION:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 60/026,062
;      FILING DATE:  13-SEP-1996
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Verny, Hana
;      REGISTRATION NUMBER:  30,518
;      REFERENCE/DOCKET NUMBER:  480.76-1(HV)
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  650-324-1677
;      TELEFAX:  650-324-1678
;  INFORMATION FOR SEQ ID NO:  6:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  1721 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
US-08-928-361B-6


  Query Match             6.7%;  Score 498.5;  DB 3;  Length 1721;
  Best Local Similarity   21.8%;  Pred. No. 5e-23;
  Matches  300;  Conservative   94;  Mismatches  510;  Indels  469;  Gaps   44;

Qy    107 STTKRSPKPPNKKKTKKVIESEEITEEHSVSENQESSSSSSSSSSSSTIWKIKS------ 160
          | |    | |||: |                :| ::        |  ||| ::: |
Db     38 SITSGELKDPNKQAT-------------ISGSRSCGWKQGYSIDSSTGFRVDSITGLPT 83

Qy    161 --SKNSAANRELQKKLKVKDNKKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQ 218
              |   |        |  |  :|:  |  |        ||    | : | | |  :|
Db     84 DPYSNCPFN-------PVTGNLVSRSTGKTIPN----TYAGVYRSN-ETKTTEPSANT-- 129

Qy    219 HNKVSTSPKITTAKP--------INPRPSLPPNSDTSKE---------------TSLTVN 255
          :  |  | :  | :|        ::|: : | ||: | |                |
Db    130 YAGVYRSNETKTTEPSANTNFLLVDPKINAPCNSENSFEQGQIFDMGSKVYIPYTKCVGV 189

Qy    256 KETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTT 315
          | ||  |  ||||   |:|                                        |||
Db    190 KHTTTTTTTTTTTTTTTTT-----------------------------------TTT 211

Qy    316 KGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPKEPAPT 375
              ||    | ||     :||    | ||  :  ||      |||  :  ||.     |
Db    212 TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 271
```

```
; Sequence 5, Application US/08700651B
; Patent No. 6015882
; GENERAL INFORMATION:
;  APPLICANT: PETERSEN, CAROLYN
;  APPLICANT: LEECH, JAMES
;  APPLICANT: NELSON, RICHARD, C.
;  APPLICANT: GUT, JIRI
;  TITLE OF INVENTION: VACCINES, ANTIBODIES, PROTEINS, GLYCOPROTEINS, DNAS AND RNAS
;  TITLE OF INVENTION: FOR PROPHYLAXIS AND TREATMENT OF Cryptosporidium parvum
;  TITLE OF INVENTION: INFECTIONS
;  FILE REFERENCE: 480.19-4(HV)
;  CURRENT APPLICATION NUMBER: US/08/700,651B
;  CURRENT FILING DATE: 1997-08-14
;  EARLIER APPLICATION NUMBER: 08/415,751
;  EARLIER FILING DATE: 1995-04-03
;  NUMBER OF SEQ ID NOS: 15
;  SOFTWARE: PatentIn Ver. 2.0
; SEQ ID NO 5
;   LENGTH: 1721
;   TYPE: PRT
;   ORGANISM: Cryptosporidium parvum
US-08-700-651-5


  Query Match             6.7%;  Score 499.5;  DB 3;  Length 1721;
  Best Local Similarity  21.8%;  Pred. No. 4.4e-23;
  Matches  300;  Conservative   94;  Mismatches  510;  Indels  469;  Gaps   44;

Qy     107 STTKRSPKPPNKKKTKKVIESEEITEEHSVSENQESSSSSSSSSSSSTIWKIKS------ 160
Db      38 SITSGERKDPNKQAT-------------ISGSRSCGWKQGYSIDSSTGFRVDSITGLPT 83

Qy     161 --SKNSAANRELQKKLKVKDNKKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQ 218
Db      84 DPYSNCPFN-------PVTGNLVSRSTGKTIPN----TYAGVYRSN-ETKTTEPSANT-- 129

Qy     219 HNKVSTSPKITTAKP--------INPRPSLPPNSDTSKE--------------TSLTVN 255
Db     130 YAGVYRSNETKTTEPSANTNFLLVDPKINAPCNSENSFEQGQIFDMGSKVYIPYTKCVGV 189

Qy     256 KETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTT 315
Db     190 KHTTTTTTTTTTTTTTTTT-----------------------------------TTT 211

Qy     316 KGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPKEPAPT 375
Db     212 TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 271

Qy     376 TTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEP 435
Db     272 TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTKKPTTTTTTTTTTTTTTTTTTTTTTT 331

Qy     436 TPTTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTP 495
Db     332 TTTTTKKPTTTTTTTTTTT-------------TTTTTTTTTTTTTTTTTTTTTTTTTT 377

Qy     496 KEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKE--------------------------- 527
Db     378 TKPTTTTTTTTTTTTTKKPTTTTTATTTTTTSETESVIKPDEWCWLEKNGECEAKGATYVG 437

Qy     528 ----------------PSPTT---------------------------------TKE 535
Db     438 VIGKDGRIENGMAFTMIPNDDTHVRFRFKVKDVGNTISVRCRKGAGKLEFPDRSLDFTIP 497

Qy     536 P--------------------APTTPKEPAPTTPKKPAPTTPKE--------------- 559
Db     498 PVAGHNSCSIIVGVSGDGKIHVSPYGSKDVSLIS----APIQPSELFNEVYCDTCTAKYG 553

Qy     560 ---PAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKP 616
Db     554 AIHSGYQTSADFVTTTTAKPTTTT--TGAPGQPTTTTTGSPSKPTTTTTTKATTTT---- 607

Qy     617 APTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTT 676
Db     608 --TTLNPIITTTTQKPTTTT------TT---KVPGKPPIATTTTTLKPIVTTTTTKATTT 656

Qy     677 PKETAPTTPKGTAPTTLKEPAPTTPKKPAPK----ELAPTTTKEPTS------------ 719
Db     657 TTTTVPTT-----TTTTKRDEMTTTTTPLPDIGDIEITPIPIEKMLDKYTRMIYDYNSGL 711

Qy     720 --TTSDKPAPTTPKG-TAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKE---PAPTTPKKP 773
Db     712 LLDSNDEPIPGSQAGQIADTSNLFPVQTHKSTGLPIDPMVGLPFDPKSGNLVHPYTNQTM 771

Qy     774 APKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEV 833
Db     772 SGLSVSYLAAKNLTVDTDETYG--LPIDTLTGYPLDPVSLIPFN--PETGELFDPISDEI 827

Qy     834 STPT--------TTKEPTTIHKS-----------------------------PDEST- 853
Db     828 MNGTIAGIVSGISASESLLSQKSALIDPATNMVVGEFGGLLNPATGVMIPGFLGPSEQTQ 887

Qy     854 ------------PELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTER--- 898
Db     888 FSPEIEDGGIIPPEVAAANADKFKLSIP--PSVP------ESIPE-----KDQKIDSISE 934

Qy     899 -------------DLRTTPETTTAAPKMTKETATTTEKTTESKITATT-TQVTSTTTQD 943
Db     935 LMYDIESGRLIGQVSKRPIPGSIAGDLNPIMKTPTQTDSVTGKPIDPTTGLPFNPPTGHL 994

Qy     944 TTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKP--EETAKPKD-------RATNSKATT 994
Db     995 INPTNNNTMDSSFAGAYKYAVSNGIKTDNVYGLPVGEITGLPKDPGSDIPFNSTTGELVD 1054

Qy     995 PKPQKPTKAPKKPTSTKKPKTMP--------------------RVRKPKTTPTPRKMTS 1033
Db    1055 PSTGKPINNSTAGIVSGKPGLPPIEDENGNLFDPSTNLPIDGNNQLVNPETNSTVSGSTS 1114

Qy    1034 --TMPE---------------------------LNPTSRIAEAMLQTTTRPNQTPNSKL 1063
Db    1115 GTTKPKPGIPVNGGGVVPDEEAKDQADKGKDGLIVPPTNSINKDPVTNTQYSNTTGNI-- 1172

Qy    1064 VEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPD--------MDYLPRVPNQGIIINP-- 1113
Db    1173 --INP--ETGKVIPGSLPGSLNYPSFNTPQQTDEITGKPVDTVTGLPYDPSTGEIIDPAT 1228

Qy    1114 ------------MLSDETNICNGK--------------PVDGLTTLRNGTLV 1139
Db    1229 KLPIPGSVAGDEILTEVLNITTDEVTGLPIDLETGLPRDPVSGLPQLPNGTLV 1281


RESULT   5
US-09-103-429A-3
; Sequence 3, Application US/09103429A
; Patent No. 6187558
;  GENERAL INFORMATION:
;    APPLICANT:  Granados, Robert R
;    APPLICANT:  Wang, Ping
;    TITLE OF INVENTION:  A No. 6187558el Invertebrate Intestinal Mucin
;    TITLE OF INVENTION:  cDNA and Related Products and Methods
;    NUMBER OF SEQUENCES:  4
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  Brown, Pinnisi & Michaels, P.C.
;      STREET:  118 No. 6187558th Tioga
;      CITY:  Ithaca
;      STATE:  NY
;      COUNTRY:  USA
;      ZIP:  14850
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/09/103,429A
;      FILING DATE:  24-JUN-1998
```



RESULT 3
5202236-25
;Patent No. 5202236
; APPLICANT: MAUGH, KATHY J.;ANDERSON, DAVID M.;STRAUSBERG,
;SUSAN L.;MCCANDLISS, RUSS;WEI, TENA;FILPULA, DAVID
; TITLE OF INVENTION: METHOD OF PRODUCING BIOADHESIVE
;PROTEIN
; NUMBER OF SEQUENCES: 39
; CURRENT APPLICATION DATA:
; APPLICATION NUMBER: US/07/528,762
; FILING DATE: 25-MAY-1990
; APPLICATION NUMBER: 82,456
; FILING DATE: 07-AUG-1987
; APPLICATION NUMBER: 933,945
; FILING DATE: 24-NOV-1986
; APPLICATION NUMBER: 650,128
; FILING DATE: 13-SEP-1984
;SEQ ID NO:25:
; LENGTH: 744
5202236-25

Query Match 6.8%; Score 506.5; DB 6; Length 744;
Best Local Similarity 29.9%; Pred. No. 6e-24;
Matches 239; Conservative 78; Mismatches 308; Indels 175; Gaps 50;



RESULT 4
US-08-700-651-5

```
  Matches  226;  Conservative   33;  Mismatches  211;  Indels  342;  Gaps   40;

Qy     300 TSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAP-----TTP 354
           |: |    |:     | |||  |    | | |     | |   |                |
Db       9 TALGLVAARPEVSDAEKNPALHEPHPDXP--PAEQXXLLPXEYDCTKFYYCEYGLKFIAP 66

Qy     355 KEPAPTT-----------TKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTK 402
           :: || |           :. | |  || ||    || | :  ||||||:  |||||
Db      67 RDCAPGTEFKFSAQTCVHAALAGCTLPGPPAETT---QAPATTQ--APTTTQ--APTTTT 119

Qy     403 SAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTA 462
            ||||  :   ||           ||||  : || |     | ||: |  ||
Db     120 QAPTTTTQATTTT----------QAPTTTQAPTTT----QATTTQAPTTTT-------- 156

Qy     463 PKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAP 522
           ||||  : ||||  : |||||:          |||||: |||||:  ||||   |
Db     157 ---QAPTTTTQ-APTTTTQ-APTTTQ----------APTTTQ-APTTTQ--APTTITQAA 198

Qy     523 TTPKEPSPTTTKEPAPTTP--KEPAPTTPKKPAPTTPKEPAPTTPK---------------- 566
           |||   :|  |  || |||    || |||  ||||:    ||  |
Db     199 TTPAATTPAAT-TPAATTPAATTPAATTPGVPAPTS----APVWPPICELLPNGCPADFD 253

Qy     567 ------------------------------------------------EPAPTTTK 574
                                                           |||  |:
Db     254 IHLLIPHDKYCNLFYQCSNGYTFEQRCPEGLYFNPYVQRCDSPANVECDGEISPAPPVTE 313

Qy     575 K----------------PA---------------------------------------- 577
                           ||
Db     314 GNEDEDIDIGDLLDNGCPANFEIDWLLPHGNRCDKYYQCVHGNLVERRCGAGTHFSFELQ 373

Qy     578 --------------------------------PTAP--KEPAPT-----------TPKET 592
                                           || |   || |             | |:
Db     374 QCDHIELVGCTLPGGESEEVDVDEDACTGWYCPTEPIEWEPLPNGCPADFSIDHLLPHES 433

Qy     593 ------------------------APTTPKKLTPTT----------PEKLAPTTPEKPAP 618
                                   :| |    :| |          :   |     ||
Db     434 DCGQYLQCVHGQTIARPCPGNLHFSPATQSCESPVTAGCQVFECDSDNQCTSTAAPTAAP 493

Qy     619 TTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPK 678
           |     |||      |:|   || | |  |||  |    ||    ||||     ||||
Db     494 TAAPTAAPTAAPTAAPSTVVPPA-TPPATAAPVPPTTAIPT----PAPTAAPTAAPTTAA 548

Qy     679 ETAPTT---PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAP 735
             :|||   |  |||     ||||    |  |:  | |  ||:  :   |||     ||
Db     549 PESPTTVTVPPTAAPTA----APTT----AVPEIPITVTSAPTAAPT--AAPTAAPTAAP 598

Qy     736 TTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTST-TSDKP 794
           ||     |||      |:|   ||||   ||| |           | |  ||:  |:  |
Db     599 TTAVPEIPTT-----VTSPPTAAPTT-AAPAPNT----------TVTVPPTAAPTTAAP 641

Qy     795 APTTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTP 854
           ||     |  | |   ||||   ||| |  | ||| :  : |     ||  |        |
Db     642 AP----NTTVTAPPTAAPTT-AAPAPNTTVTVPPTAAPTAAP-----PTVAH------AP 685

Qy     855 ELSAEPTPKALENSPKEPGVPTTKTPAATKPE 886
             :| |             | ||  || | ||
Db     686 NTTAAP------------VTTTSAPATT-PE 703


RESULT   2
US-08-928-361B-5
; Sequence 5, Application US/08928361B
; Patent No. 6071518
;  GENERAL INFORMATION:
;    APPLICANT:  Petersen, Carolyn
;    TITLE OF INVENTION:  PEPTIDES, POLYPEPTIDES, GLYCOPROTEINS,
;    TITLE OF INVENTION:  THEIR FUNCTIONAL MUTANTS, VARIANTS, ANALOGS AND FRAGMENTS
;    TITLE OF INVENTION:  FOR TREATMENT AND DETECTION/DIAGNOSIS OF CRYPTOSPORIDIUM
;    TITLE OF INVENTION:  SPECIES INFECTIONS
;    NUMBER OF SEQUENCES:  30
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  PETERS, VERNY, JONES & BIKSA
;      STREET:  385 Sherman Avenue, Suite 6
;      CITY:  Palo Alto
;      STATE:  CA
;      COUNTRY:  USA
;      ZIP:  94306-1840
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/08/928,361B
;      FILING DATE:  12-SEP-1997
;      CLASSIFICATION:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 60/026,062
;      FILING DATE:  13-SEP-1996
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Verny, Hana
;      REGISTRATION NUMBER:  30,518
;      REFERENCE/DOCKET NUMBER:  480.76-1(HV)
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  650-324-1677
;      TELEFAX:  650-324-1678
;  INFORMATION FOR SEQ ID NO:  5:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  1837 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
US-08-928-361B-5


  Query Match             6.9%;  Score 513;  DB 3;  Length 1837;
  Best Local Similarity   21.6%;  Pred. No. 6.8e-24;
  Matches  307;  Conservative  105;  Mismatches  539;  Indels  470;  Gaps   49;

Qy      84 HNPTSPPSSKKAPP----PSGASQTIKSTTKRSPKP--------------PNKKKTKKV 124
           |   |  ||  : |     ||    |:  |:|   |                 ||   | :
Db      83 HADASTTSSAYSAPFELDVSGV--PIEPNTRRMVDPVSLMLFDNSTGVMYDPN---TNSI 137

Qy     125 IESE--EITEEHSVSENQESSSSSSSSSSSSTIWKIKSSKNSAANRELQKKLKVKDNKKN 182
           :|      | |  :      :|::  :: :|  | :  |  |    ::  |:  :  ::
Db     138 LEGSIAGIRSESCIVSELNFTSTTGFTTDTSMNWPV--SITSGELKDPNKQATISGSRSC 195

Qy     183 RTKKKPTPKPPVVDEAGSGLDNG-DFKVTTPDTSTTQHNKVSTSPKITTAKPINPRPSLP 241
           |:            |  :|:   |:| :                   |  | :| |: |
Db     196 GWKQ-----------GYSIDSSTGFRVDS----------------ITGLPTDPYPNCP 226

Qy     242 PNSDTSKETSLTVNKET------TVETKETTTTNKQTST-----DGK-----EKTTSAKE 285
           | |    | :  |           : || ||     :|     | |          | ::
Db     227 FNPVTGNLVSRSTGKTIPNTYAGVYRSNETKTTEPSANTNFLLVDPKINAPCNSENSFEQ 286

Qy     286 TQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPT 345
           |  :  |   :  |  |  | |     |||     ||    | ||     :||    | |
Db     287 VQIFDMGSKVYIPYTKCVGVKHTTTTTTT-----TTTTTTTTTTTTTTTTTTTTTTTTT 341

Qy     346 TIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAP 405
           |  :  ||      |||  :  ||      |||      || |:|  |||      ||| :
Db     342 TTTTTTTT------TTTTTTTTTTTTTTTTTTTTTTTTTTKKPTTTTTTTTTTTTTTTTT 395

Qy     406 TTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKK 465
           ||       ||       ||       || |  | ||       ||      ||      | ||
Db     396 TTTTTTTTTTTTTTTTTTTTTTTTTTKKPTTTTTTTTTTTTTTTTTTTTTTTTTTTTKK 455

Qy     466 PAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTP 525
           |  ||      ||      |||   : ||      ||| :| ||||:|  |||  :  ||
Db     456 PTTTT------TTTTTTTTTTTTTTTTTTTTTTTTTTTTATTTTKKPTTTTTTTTTTTT 509

Qy     526 KEPSPTTTKEPAPTTPKEPAPTTPKK---------------------------------- 551
           |:|: |||      || :  :   | :
Db     510 KKPTTTTTATTTTTTSETESVIKPDEWCWLEKNGECEAKGATYVGVIGKDGRIENGMAFT 569
```

GenCore version 4.5
Copyright (c) 1993 - 2000 Compugen Ltd.

OM protein - protein search, using sw model

Run on: April 26, 2002, 16:16:35 ; Search time 49.78 Seconds
(without alignments)
623.836 Million cell updates/sec

Title: US-09-556-246-1_COPY_25_1404
Perfect score: 7410
Sequence: 1 QDLSSCAGRCGEGYSRDATC.....ARAITTRSGQTLSKVWYNCP 1380

Scoring table: BLOSUM62
Gapop 10.0 , Gapext 0.5

Searched: 212252 seqs, 22503292 residues

Total number of hits satisfying chosen parameters: 212252

Minimum DB seq length: 0
Maximum DB seq length: 2000000000

Post-processing: Minimum Match 0%
Maximum Match 100%
Listing first 45 summaries

Database : Issued_Patents_AA:*
1: /cgn2_6/ptodata/2/iaa/5A_COMB.pep:*
2: /cgn2_6/ptodata/2/iaa/5B_COMB.pep:*
3: /cgn2_6/ptodata/2/iaa/6A_COMB.pep:*
4: /cgn2_6/ptodata/2/iaa/6B_COMB.pep:*
5: /cgn2_6/ptodata/2/iaa/PCTUS_COMB.pep:*
6: /cgn2_6/ptodata/2/iaa/backfiles1.pep:*

Pred. No. is the number of results predicted by chance to have a score greater than or equal to the score of the result being printed, and is derived by analysis of the total score distribution.

SUMMARIES

Result No.	Score	% Query Match	Length	DB	ID	Description
1	522	7.0	805	4	US-09-103-429A-4	Sequence 4, Appli
2	513	6.9	1837	3	US-08-928-361B-5	Sequence 5, Appli
3	506.5	6.8	744	6	5202236-25	Patent No. 5202236
4	499.5	6.7	1721	3	US-08-700-651-5	Sequence 5, Appli
5	498.5	6.7	786	4	US-09-103-429A-3	Sequence 3, Appli
6	498.5	6.7	1721	3	US-08-928-361B-6	Sequence 6, Appli
7	488.5	6.6	826	1	US-07-638-431-2	Sequence 2, Appli
8	488.5	6.6	826	5	PCT-US92-00018-2	Sequence 2, Appli
9	476.5	6.4	652	6	5202236-13	Patent No. 5202236
10	452	6.1	960	4	US-09-219-849-5	Sequence 5, Appli
11	425.5	5.7	1185	4	US-09-041-886-23	Sequence 23, Appl
12	424.5	5.7	1867	2	US-08-479-537A-5	Sequence 5, Appli
13	424.5	5.7	1867	4	US-09-083-116-5	Sequence 5, Appli
14	424.5	5.7	2035	2	US-08-479-537A-2	Sequence 2, Appli
15	424.5	5.7	2035	4	US-09-083-116-2	Sequence 2, Appli
16	419.5	5.7	2476	2	US-08-276-967-2	Sequence 2, Appli
17	417	5.6	829	1	US-08-642-255-132	Sequence 132, App
18	417	5.6	829	1	US-08-397-633A-53	Sequence 53, Appl
19	417	5.6	837	1	US-08-175-155-68	Sequence 68, Appl
20	417	5.6	837	1	US-08-477-509B-103	Sequence 103, App
21	417	5.6	837	1	US-08-642-255-101	Sequence 101, App
22	417	5.6	837	2	US-08-707-237A-75	Sequence 75, Appl
23	417	5.6	837	3	US-08-482-085B-103	Sequence 103, App
24	417	5.6	897	1	US-08-397-633A-50	Sequence 50, Appl
25	413.5	5.6	907	3	US-08-783-774-2	Sequence 2, Appli
26	413.5	5.6	907	5	PCT-US95-04611A-19	Sequence 19, Appl
27	404.5	5.5	408	1	US-07-609-716-65	Sequence 65, Appl
28	404.5	5.5	408	4	US-08-475-411A-65	Sequence 65, Appl
29	404.5	5.5	408	4	US-08-478-029A-65	Sequence 126, App
30	398.5	5.4	682	1	US-08-642-255-126	Sequence 36, Appl
31	398.5	5.4	682	1	US-08-397-633A-36	Sequence 2, Appli
32	384	5.2	1537	1	US-08-325-267A-2	Sequence 4, Appli
33	383.5	5.2	1231	3	US-08-904-263A-4	Sequence 6, Appli
34	381	5.1	1848	4	US-08-296-791-6	Sequence 6, Appli
35	381	5.1	1848	5	PCT-US95-10661A-6	Sequence 6, Appli
36	376	5.1	960	4	US-09-219-849-6	Sequence 8, Appli
37	373	5.0	1786	4	US-08-973-462-8	Sequence 84, Appl
38	371	5.0	761	2	US-08-707-237A-84	Sequence 114, App
39	371	5.0	762	1	US-08-642-255-114	Sequence 26, Appl
40	371	5.0	762	1	US-08-397-633A-26	Sequence 62, Appl
41	370.5	5.0	1064	1	US-08-642-255-62	Sequence 28, Appl
42	370.5	5.0	1187	1	US-08-320-559-28	Sequence 28, Appl
43	370.5	5.0	1187	3	US-08-545-860D-28	Sequence 28, Appl
44	370.5	5.0	1187	5	PCT-US94-04496-28	Sequence 26, Appl
45	370.5	5.0	1210	1	US-08-320-559-26	

ALIGNMENTS

RESULT 1
US-09-103-429A-4
; Sequence 4, Application US/09103429A
; Patent No. 6187558
; GENERAL INFORMATION:
; APPLICANT: Granados, Robert R
; APPLICANT: Wang, Ping
; TITLE OF INVENTION: A No. 6187558el Invertebrate Intestinal Mucin
; TITLE OF INVENTION: cDNA and Related Products and Methods
; NUMBER OF SEQUENCES: 4
; CORRESPONDENCE ADDRESS:
; ADDRESSEE: Brown, Pinnisi & Michaels, P.C.
; STREET: 118 No. 6187558th Tioga
; CITY: Ithaca
; STATE: NY
; COUNTRY: USA
; ZIP: 14850
; COMPUTER READABLE FORM:
; MEDIUM TYPE: Floppy disk
; COMPUTER: IBM PC compatible
; OPERATING SYSTEM: PC-DOS/MS-DOS
; SOFTWARE: PatentIn Release #1.0, Version #1.30
; CURRENT APPLICATION DATA:
; APPLICATION NUMBER: US/09/103,429A
; FILING DATE: 24-JUN-1998
; CLASSIFICATION: 800
; ATTORNEY/AGENT INFORMATION:
; NAME: Michaels, Christopher A
; REGISTRATION NUMBER: 34,390
; REFERENCE/DOCKET NUMBER: BTI-39
; TELECOMMUNICATION INFORMATION:
; TELEPHONE: (607) 256-2000
; TELEFAX: (607) 256-3628
; INFORMATION FOR SEQ ID NO: 4:
; SEQUENCE CHARACTERISTICS:
; LENGTH: 805 amino acids
; TYPE: amino acid
; STRANDEDNESS: single
; TOPOLOGY: linear
; MOLECULE TYPE: protein
; HYPOTHETICAL: NO
; ANTI-SENSE: NO
; ORIGINAL SOURCE:
; ORGANISM: Trichoplusia ni
; TISSUE TYPE: peritrophic membrane
US-09-103-429A-4

Query Match 7.0%; Score 522; DB 4; Length 805;
Best Local Similarity 27.8%; Pred. No. 7.2e-25;

```
DR   PIR; S00979; QFHUH.
DR   MIM; 162230; -.
DR   InterPro; IPR001664; IF.
DR   Pfam; PF00038; filament; 1.
DR   PROSITE; PS00226; IF; 1.
KW   Intermediate filament; Coiled coil; Neurone; Phosphorylation.
FT   DOMAIN       1    100       HEAD.
FT   DOMAIN     101    413       ROD.
FT   DOMAIN     414   1020       TAIL.
FT   DOMAIN     101    132       COIL 1A.
FT   DOMAIN     133    145       LINKER 1.
FT   DOMAIN     146    244       COIL 1B.
FT   DOMAIN     245    266       LINKER 12.
FT   DOMAIN     267    288       COIL 2A.
FT   DOMAIN     289    292       LINKER 2.
FT   DOMAIN     293    413       COIL 2B.
SQ   SEQUENCE   1020 AA;  111780 MW;  1177C9DCB3DCF1D4 CRC64;
```

Query Match 6.8%; Score 502; DB 1; Length 1020;
Best Local Similarity 26.7%; Pred. No. 3.4e-13;
Matches 209; Conservative 81; Mismatches 356; Indels 136; Gaps 35;

```
Qy    119 KKTKKVIESE--EITEEHSVS-ENQESSSSSSSSSSSSTIWKIKSSKNSAANRELQKKLK 175
Db    336 KSTKDSLERQRSELEDRHQADIASYQEAIQQLDAELRNTKWEM-----AAQLREYQDLLN 390

Qy    176 VK-------------DNKKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNK 221
Db    391 VKMALDIEIAAYRKLLEGEECRIGFGPIP---------FSLPEGLPKI--PSVST--HIK 437

Qy    222 VSTSPKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKE-KT 280
Db    438 VKSEEKIKVVE------------KSEKETVIVEE-----QTEETQVTEEVTEEEEKEAKE 480

Qy    281 TSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPK 340
Db    481 EEGKEEEGGEEEEAEG-------------GEEETKSP----PAEEAASPEKEAKSPVKE 522

Qy    341 EPTPTTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTT 400
Db    523 EAKSPAEAKSPEKEEAKSPAEVKS-PEKAKSPA----KEEAKSPPEAKSPEKEEAKSPAE 577

Qy    401 TKSAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAP 460
Db    578 VKS-PEKAKSPAKEEAKSPA----EAKSPEKAKSPVKEEAKSPA-EAKSPVKEEAKSPAE 631

Qy    461 TAPKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKS 520
Db    632 VKSPEKAKSPTKEEAKSPEKAKSPEKEEAKSPEKAKSPVKAEAKS-PEKAKSPVKAEAKS 690

Qy    521 APTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTA 580
Db    691 -PEKAKSP----VKEEAKSPEKAKSPVKEEAKSPEKAKSPV----KEEAKTPEKAKSPV- 740

Qy    581 PKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPE--KPAPTTPEEL-APTTPEEPTPTTP 637
Db    741 -KEEAKSPEKAKSPEKAKTLDVKSPEAKTPAKEEARSPADKFPEKAKSPVKEEVKSPEKA 799

Qy    638 EEPAPTTPKAAAPNTPKEPAPTTP----KEPAPTTPKEPAPTTPKETAPTTPKGTAPTTL 693
Db    800 KSPLKADAKAPEKEIPKKEEVKSPVKEEEKPQEVKVKEPPKKAEEEKAPATPK------- 852

Qy    694 KEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTT 753
Db    853 TEEKKDSKKEEAPKKEAP----KPKVEEKKEPAVEKPKESKVEAKKEEAEDKKKVP---T 905

Qy    754 PKGTAP--TTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPA-PTTPKETAPTTPKEP 810
Db    906 PEKEAPAKVEVKEDA-------KPKEKTEVAKKEPDDAKAKEPSKPAEKKEAAP-EKKDT 957

Qy    811 APTTPKKPAPTTPETPPPTTSEV--STPTTTKEPTTIHKSPDE---STPELSAEPTPKAL 865
```

```
Db    958 KEEKAKK-----PEEKPKTEAKAKEDDKTLSKEPSKPKAEKAEKSSSTDQKDSKPPEKAT 1012

Qy    866 EN 867
Db   1013 ED 1014
```

Search completed: April 26, 2002, 16:19:21
Job time: 346 sec

```
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=95347616; PubMed=7622068;
RA   Mita K., Tsuji H., Morimyo M., Takahashi E., Nenoi M.,
RA   Ichimura S., Yamauchi M., Hongo E., Hayashi A.;
RT   "The human gene encoding the largest subunit of RNA polymerase II.";
RL   Gene 159:285-286(1995).
CC   -!- FUNCTION: DNA-DEPENDENT RNA POLYMERASE CATALYZES THE TRANSCRIPTION
CC       OF DNA INTO RNA USING THE FOUR RIBONUCLEOSIDE TRIPHOSPHATES AS
CC       SUBSTRATES.
CC   -!- CATALYTIC ACTIVITY: N NUCLEOSIDE TRIPHOSPHATE = N PYROPHOSPHATE +
CC       RNA(N).
CC   -!- SUBUNIT: RNA POLYMERASE II CONSISTS OF 12 DIFFERENT SUBUNITS.
CC   -!- SUBCELLULAR LOCATION: NUCLEAR.
CC   -!- PTM: THE TANDEM 7 RESIDUES REPEATS CAN BE HIGHLY PHOSPHORYLATED.
CC       THE PHOSPHORYLATION ACTIVATES POL2.
CC   -!- MISCELLANEOUS: THREE DISTINCT ZINC-CONTAINING RNA POLYMERASES ARE
CC       FOUND IN EUKARYOTIC NUCLEI: POLYMERASE I FOR THE RIBOSOMAL RNA
CC       PRECURSOR, POLYMERASE II FOR THE MRNA PRECURSOR, AND POLYMERASE
CC       III FOR 5S AND TRNA GENES.
CC   -!- SIMILARITY: BELONGS TO THE RNA POLYMERASE BETA' CHAIN FAMILY.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X63564; CAA45125.1; -.
DR   EMBL; X74874; CAA52862.1; -.
DR   EMBL; X74873; CAA52862.1; JOINED.
DR   EMBL; X74872; CAA52862.1; JOINED.
DR   EMBL; X74871; CAA52862.1; JOINED.
DR   EMBL; X74870; CAA52862.1; JOINED.
DR   PIR; S21054; S21054.
DR   MIM; 180660; -.
DR   InterPro; IPR000684; RNA_polII_repeat.
DR   InterPro; IPR000722; RNA_pol_A.
DR   InterPro; IPR002879; RNA_pol_A2.
DR   Pfam; PF00623; RNA_pol_A; 1.
DR   Pfam; PF01854; RNA_pol_A2; 1.
DR   PROSITE; PS00115; RNA_POL_II_REPEAT; 43.
KW   Transferase; DNA-directed RNA polymerase; Transcription; Zinc; Repeat;
KW   DNA-binding; Nuclear protein; Phosphorylation; Zinc-finger.
FT   ZN_FING      71     87       C2H2-TYPE (POTENTIAL).
FT   DOMAIN     1590   1958       CARBOXYL-TERMINAL 7-RESIDUE REPEATS.
FT   CONFLICT   1067   1067       W -> L (IN REF. 2).
FT   CONFLICT   1449   1449       D -> Y (IN REF. 2).
SQ   SEQUENCE   1970 AA;  217205 MW;  6876FC25692A657E CRC64;

  Query Match             6.8%;  Score 503.5;  DB 1;  Length 1970;
  Best Local Similarity   33.3%;  Pred. No. 5.1e-13;
  Matches  181;  Conservative  101;  Mismatches  167;  Indels   95;  Gaps   47;

Qy     326 PTPTTPKEPASTTPKE-PTPTTIKSAPTTPKEPAPTTTKSAPTTPKE-----PAPTTTKE 379
           |:|     || |   :  ||     :|:       |      :|:   :     |  :
Db    1507 PSPMGGISPAMTPWNQGATPAYGAWSPSVGSGMTPGAAGFSPSAASDASGFSPGYSPAWS 1566

Qy     380 PAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEP--APTTPKEPTP 437
           | | :|  | |::   |:|    |   :|  ||  -|: |   ||: |  :||:|   :|
Db    1567 PTPGSPGSPGPSSPYIPSPGGAMSPSYSPTSPA-YEPRSPGGYTPQSPSYSPTSPSY-SP 1624

Qy     438 TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKE 497
           |:|   :||:   :||:|   :||:|   :||:|   :||:|           |||:|
Db    1625 TSPSY-SPTSPNYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY---------SPTSPSY 1671

Qy     498 PAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTP 557
           :||:   :||:   :||:   :||:|   |||:   :||:|   :||:|   :||:|
Db    1672 -SPTSPSYSPTSPSY-SPTSPSYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY-SPTSP 1725

Qy     558 KEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPA 617
           :||:|   :||:    :||:|   :||:|  | ||:|   :||:|   :||:|
Db    1726 SY-SPTSPSY-SPTSPSY-SPTSPNY-SPTSPNYT-PTSP-SYSPTSP-SYSPTSPNY-T 1777

Qy     618 PTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEP--APT 675
           ||:|   :||:|   :||:|   :||:|  : :|:::|      ||: |   ||  |  :|:
Db    1778 PTSP-NYSPTSPSY-SPTSPSY-SPTSP-SYSPSSPR----YTPQSPT-YTPSSPSYSPS 1828

Qy     676 TPKETAPTTPKGTAPTTLKEPAPTTPK-KPAPKELAPTTTKEPTSTTSDKPAPTTPKGTA 734
           :|   :||:|| |          ||:|   |:  |  ||: |   | || | :|||:||  :
Db    1829 SP-SYSPTSPKYT---------PTSPSYSPSSPEYTPTSPK--YSPTSPKYSPTSPK-YS 1875

Qy     735 PTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPT-STTSDK 793
           ||:|   :|||||  :||:|  |   |     ||:||       :||:   || | || |
Db    1876 PTSPTY-SPTTPKY-SPTSP--TYSPTSPVYTPTSPK------YSPTS---PTYSPTSPK 1922

Qy     794 PAPTTPKETAPTTPKEP--APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDE 851
           :||:|   :||:||    :||:|   :||:|      | :::|              |||:
Db    1923 YSPTSP-TYSPTSPKGSTYSPTSPGY-SPTSP------TYSLTSPAI---------SPDD 1965

Qy     852 STPE 855
           |  |
Db    1966 SDEE 1969


RESULT  15
NFH_HUMAN
ID   NFH_HUMAN      STANDARD;      PRT;  1020 AA.
AC   P12036;
DT   01-OCT-1989 (Rel. 12, Created)
DT   01-OCT-1989 (Rel. 12, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   NEUROFILAMENT TRIPLET H PROTEIN (200 KDA NEUROFILAMENT PROTEIN)
DE   (NEUROFILAMENT HEAVY POLYPEPTIDE) (NF-H).
GN   NEFH OR NFH.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=88328981; PubMed=3138108;
RA   Lees J.F., Shneidman P.S., Skuntz S.F., Carden M.J., Lazzarini R.A.;
RT   "The structure and organization of the human heavy neurofilament
RT   subunit (NF-H) and the gene encoding it.";
RL   EMBO J. 7:1947-1955(1988).
CC   -!- FUNCTION: NEUROFILAMENTS USUALLY CONTAIN THREE IF PROTEINS: L, M,
CC       AND H WHICH ARE INVOLVED IN THE MAINTENANCE OF NEURONAL CALIBER.
CC       NF-H HAS AN IMPORTANT FUNCTION IN MATURE AXONS THAT IS NOT
CC       SUBSERVED BY THE TWO SMALLER NF PROTEINS.
CC   -!- PTM: THERE ARE A NUMBER OF REPEATS OF THE TRIPEPTIDE K-S-P, NFH IS
CC       PHOSPHORYLATED ON A NUMBER OF THE SERINES IN THIS MOTIF. IT IS
CC       THOUGHT THAT PHOSPHORYLATION OF NFH RESULTS IN THE FORMATION OF
CC       INTERFILAMENT CROSS BRIDGES THAT ARE IMPORTANT IN THE MAINTENANCE
CC       OF AXONAL CALIBER.
CC   -!- PTM: PHOSPHORYLATION SEEMS TO PLAY A MAJOR ROLE IN THE FUNCTIONING
CC       OF THE LARGER NEUROFILAMENT POLYPEPTIDES (NF-M AND NF-H), THE
CC       LEVELS OF PHOSPHORYLATION BEING ALTERED DEVELOPMENTALLY AND
CC       COINCIDENT WITH A CHANGE IN THE NEUROFILAMENT FUNCTION.
CC   -!- SIMILARITY: BELONGS TO THE INTERMEDIATE FILAMENT FAMILY.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X15306; CAA33366.1; -.
DR   EMBL; X15307; CAA33366.1; JOINED.
DR   EMBL; X15308; CAA33366.1; JOINED.
DR   EMBL; X15309; CAA33366.1; JOINED.
```

```
DR   PROSITE; PS00025; P_TREFOIL; 6.
KW   Repeat; Amphibian skin; Glycoprotein; Alternative splicing.
FT   NON_TER       1      1
FT   DOMAIN       81    144       8 X 8 AA APPROXIMATE TANDEM REPEATS,
FT                                ALA/THR-RICH.
FT   REPEAT       81     88       1-1.
FT   REPEAT       89     96       1-2.
FT   REPEAT       97    104       1-3.
FT   REPEAT      105    112       1-4.
FT   REPEAT      113    120       1-5.
FT   REPEAT      121    128       1-6.
FT   REPEAT      129    136       1-7.
FT   REPEAT      137    144       1-8.
FT   DOMAIN      161    202       P-TYPE 1.
FT   DOMAIN      218    301       8 X APPROXIMATE TANDEM REPEATS, THR-RICH.
FT   REPEAT      218    224       2-1.
FT   REPEAT      225    239       2-2.
FT   REPEAT      240    249       2-3.
FT   REPEAT      250    259       2-4.
FT   REPEAT      260    275       2-5.
FT   REPEAT      276    287       2-6.
FT   REPEAT      288    294       2-7.
FT   REPEAT      295    301       2-8.
FT   DOMAIN      306    347       P-TYPE 2.
FT   DOMAIN      353    394       P-TYPE 3.
FT   DOMAIN      402    522       12 X APPROXIMATE TANDEM REPEATS,
FT                                THR-RICH.
FT   REPEAT      402    411       3-1.
FT   REPEAT      412    419       3-2.
FT   REPEAT      420    431       3-3.
FT   REPEAT      432    443       3-4.
FT   REPEAT      444    453       3-5.
FT   REPEAT      454    460       3-6.
FT   REPEAT      461    472       3-7.
FT   REPEAT      473    479       3-8.
FT   REPEAT      480    491       3-9.
FT   REPEAT      492    498       3-10.
FT   REPEAT      499    515       3-11.
FT   REPEAT      516    522       3-12.
FT   DOMAIN      525    566       P-TYPE 4.
FT   DOMAIN      572    613       P-TYPE 5.
FT   DOMAIN      620    661       P-TYPE 6.
FT   DISULFID    162    188       BY SIMILARITY.
FT   DISULFID    172    187       BY SIMILARITY.
FT   DISULFID    182    199       BY SIMILARITY.
FT   DISULFID    307    333       BY SIMILARITY.
FT   DISULFID    317    332       BY SIMILARITY.
FT   DISULFID    327    344       BY SIMILARITY.
FT   DISULFID    354    380       BY SIMILARITY.
FT   DISULFID    364    379       BY SIMILARITY.
FT   DISULFID    374    391       BY SIMILARITY.
FT   DISULFID    526    552       BY SIMILARITY.
FT   DISULFID    536    551       BY SIMILARITY.
FT   DISULFID    546    563       BY SIMILARITY.
FT   DISULFID    573    599       BY SIMILARITY.
FT   DISULFID    583    598       BY SIMILARITY.
FT   DISULFID    593    610       BY SIMILARITY.
FT   DISULFID    621    647       BY SIMILARITY.
FT   DISULFID    631    646       BY SIMILARITY.
FT   DISULFID    641    658       BY SIMILARITY.
FT   VARIANT     276    276       K -> E.
FT   VARIANT     354    354       C -> R.
FT   VARIANT     415    415       T -> A.
SQ   SEQUENCE   662 AA;  67774 MW;  F085277F1ED2FD40 CRC64;

  Query Match             6.9%;  Score 509.5;  DB 1;  Length 662;
  Best Local Similarity   28.1%;  Pred. No. 1.2e-13;
  Matches  223;  Conservative   48;  Mismatches  276;  Indels  247;  Gaps   30;

Qy     267 TTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEP 326
           ||    :  ||: | :|: : : |||:|       :| | ||  |     |   ||
Db       3 TTAAAVAATGKDTTAAAEGSAAAEKTAA-----AGEVSAPPT--AAVAATGEDATT---- 51

Qy     327 TPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPK 386
               |: |    |      ||||   || |    ||||    ||||    ||||
Db      52 --------AAATAAAETTAAAGEAPTTTAPATTAAGKAPTTAAATAPTTAAAGAPTTAT 103

Qy     387 EPAPTTTKEPAPTTTKS-APTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPA- 444
           || |   | |||  | ||||      :|    ||||    | :  :  : :: :|
Db     104 GKAPATAAAPVPTTAASKAPTTAAAATHSTAAAAAPTTAASAAKSKERSTSSSSEEEHCH 163

Qy     445 --PTTKEPAPT---TPKEPAPTAPKKPAPTTPK----------EPAPTTPKEPAPTTTKE 489
             |: :|   :   | |:      ||     ||          :|   : :|   ||||
Db     164 VKPSKREMCGSKGITKKQ----CKKKNCCFDPKGHGGIHCFHRKPKGHSHEEHTTTTTK- 218

Qy     490 PSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTP 549
               |||| : | |||    ||||             ||| :  |||
Db     219 ---------APTTIQIATTTT---TPTTT-----------TTTTKATPTT--------- 245

Qy     550 KKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLA 609
                   ||  :  ||||                    || | |  ||    ||||
Db     246 ------------TTTTKATPTTT-----------TTTKATTTTT----TPTT----- 270

Qy     610 PTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTP 669
                 ||    | |||   | |||
Db     271 ---------TTTTTKATTTPTTTTTTTP------------------------------ 289

Qy     670 KEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTT 729
                   |  ||   |  ||       | |          :|   :   :
Db     290 ----------TTTTTKATTTTTTTSGECKMEPSK----------REDCGYSGITESQCR 328

Qy     730 PKG-----TAPTT-----PKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELA 779
           ||     : | |             ||:     |    |    |    :|    : :
Db     329 TKGCCFDSSIPQTKWCFYTLSQVADCKVEPSQRVDCGFRGIT----ADQCRQKNCCFDSS 384

Qy     780 PTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTT 839
           : ||    :||   |  |   | |||   |  ||  |   |||      |:  :|||||
Db     385 ISGTKWCFYSTSQVAA--TKTTTPTTTTTPTTTTTTKATTTP------TTTTTPTTT 436

Qy     840 KEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERD 899
             ||                | ||                || ||  | |  |||   | |
Db     437 TTTTT---------------TTKA----------TTTTPTTTTPTTTTT---KAT--- 463

Qy     900 LRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAP 959
             ||| |||  |     | |||: |       |||  |:||| ||  | ||   ||
Db     464 -TTTPTTTTTTP-----TTTTTKAT-------TTTPTTTTTTPTTTTTKATTTTPTTTTT 510

Qy     960 KVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPK----T 1015
             |||| | |||    | |    |  ||        : |  :|     :|  : |   :
Db     511 TTTTTKATTTTTSGECKME----PSKRADCGYPGITESQCRSKGCCFDSSIPQTKWCFYS 566

Qy    1016 MPRVRKPKTTPTPR 1029
           :|:|   |  |: |
Db     567 LPQVADCKVAPSSR 580


RESULT  14
RPB1_HUMAN
ID   RPB1_HUMAN     STANDARD;      PRT;  1970 AA.
AC   P24928;
DT   01-MAR-1992 (Rel. 21, Created)
DT   01-MAR-1992 (Rel. 21, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   DNA-DIRECTED RNA POLYMERASE II LARGEST SUBUNIT (EC 2.7.7.6) (RPB1).
GN   POLR2A.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=92178992; PubMed=1542581;
RA   Wintzerith M., Acker J., Vicaire S., Vigneron M., Kedinger C.;
RT   "Complete sequence of the human RNA polymerase II largest subunit.";
RL   Nucleic Acids Res. 20:910-910(1992).
```

```
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; M37227; AAA41693.1; ALT_FRAME.
DR   EMBL; X13804; CAA32038.1; ALT_FRAME.
DR   EMBL; M21964; AAA41695.1; -.
DR   EMBL; J04517; AAA41692.1; -.
DR   PIR; A30796; A30796.
DR   PIR; A25649; A25649.
DR   PIR; B25649; B25649.
DR   PIR; S02003; S02003.
DR   InterPro; IPR001664; IF.
DR   Pfam; PF00038; filament; 1.
DR   PROSITE; PS00226; IF; 1.
KW   Intermediate filament; Coiled coil; Neurone; Phosphorylation; Repeat.
FT   NON_TER       1      1
FT   DOMAIN      276    641       51 X 3 AA TANDEM REPEATS OF K-S-P.
FT   CONFLICT    164    164       L -> I (IN REF. 2).
FT   CONFLICT    185    185       I -> S (IN REF. 2).
FT   CONFLICT    193    193       L -> T (IN REF. 2).
FT   CONFLICT    199    199       M -> T (IN REF. 2).
FT   CONFLICT    346    346       K -> N (IN REF. 1).
FT   CONFLICT    373    373       A -> V (IN REF. 1 AND 4).
FT   CONFLICT    482    482       G -> E (IN REF. 2 AND 4).
FT   CONFLICT    485    485       P -> S (IN REF. 2).
FT   CONFLICT    570    571       RK -> KE (IN REF. 2 AND 4).
FT   CONFLICT    591    591       P -> T (IN REF. 2 AND 4).
FT   CONFLICT    727    727       A -> V (IN REF. 4).
FT   CONFLICT    757    759       AAP -> GST (IN REF. 4).
FT   CONFLICT    769    769       T -> L (IN REF. 2).
FT   CONFLICT    775    775       R -> P (IN REF. 2 AND 4).
SQ   SEQUENCE   831 AA;  89486 MW;  1B0973C3F13EF768 CRC64;


  Query Match             7.0%;  Score 517.5;  DB 1;  Length 831;
  Best Local Similarity   26.6%;  Pred. No. 7.2e-14;
  Matches  217;  Conservative   76;  Mismatches  345;  Indels  179;  Gaps   41;

Qy     119 KKTKKVIESE--EITEEHSVS-ENQESSSSSSSSSSSSTIWKIKSSKNSAANRELQKKLK 175
           | ||: :| :  |: : | |    : : :     :   :| |::     :|  || |  |
Db     103 KSTKESLERQRSELEDRHQVDMASYQDAIQQLDNELRNTKWEM-----AAQLREYQDLLN 157

Qy     176 VK--------------DNKKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNK 221
           ||              : :: |    | |           |  |  |:  |  ||  | |
Db     158 VKMALDLEIAAYRKLLEGEECRIGFGPIP---------FSLTEGLPKI--PSMST--HIK 204

Qy     222 VSTSPKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTT 281
           | :  ||   :             : ||| :   :   ::  |  |  :      |:|:
Db     205 VKSEEKIKVVE-----------KSEKETVIVEEQTEEIQVTEEVTEEEDKEAQGEEEEE 252

Qy     282 SAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEP----TPTTPKEPA-S 336
           : :  :    ||     | ::  | |    |  || |     | |    :|  | || :
Db     253 AEEGGEEAATTS-----PPAEEAASP----EKETKSPVKEEAKSPAEAKSPAEAKSPAEA 303

Qy     337 TTPKEPTPTTIKSAPTTPKEPA----PTTTKSAPTTPKEPA----PTTTKEP----APTT 384
           :| |     :  :|   | ||    |   || |   | ||    | | | |    :|
Db     304 KSPAEVKSPAVAKSPAEVKSPAEVKSPAEAKS-PAEAKSPAEVKSPATVKSPGEAKSPAE 362

Qy     385 PKEPA----PTTTKEPA----PTTTKSAPTTPKEPA----PTTPKKPA----PTTPKEPA 428
           | ||    |   | ||    | : || |   | ||    |   | ||     |   | ||
Db     363 AKSPAEVKSPAEAKSPAEAKSPASVKS-PGEAKSPAEAKSPAEVKSPATVKSPVEAKSPA 421

Qy     429 ----PTTPKEP----TPTTPKEPAPTTKEP----APTTPKEPA----PTAPKKPAPTTPK 472
               | | | |    :|   | || : | |    :|   | ||    |   | ||    |
Db     422 EVKSPVTVKSPAEAKSPVEVKSPA-SVKSPSEAKSPAGAKSPAEAKSPVVAKSPAEA--K 478

Qy     473 EPAPTTPKEPAPTTTKEPSPTTPKEPA----PTTTKS-----APTTTKEPAPTTTKS--- 520
```

```
           ||   |   | :  :  ||   | ||    |   ||     :|   | |     ||
Db     479 SPAGAKPPAEAKSPAEAKSPAEAKSPAEAKSPAEAKSPVEVKSPEKAKSPVKEGAKSLAE 538

Qy     521 --APTTPKEPSPTTTKEPA----PTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTK 574
             :|   | |     | ||    |   | |     | |       :      | ||    |
Db     539 AKSPEKAKSPVKEEIKPPAEVKSPEKAKSPMRKEAKSPEKAKTLDVKSPEAKPPAKEEAK 598

Qy     575 KPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTP 634
           :||     |   :  || | :  |: | |  ||:||    |  :|   ||:     :||
Db     599 RPADIRSPEQVKSPAKEEAKSPEKEETRT--EKVAPKKEEVKSPV--EEV---KAKEPPK 651

Qy     635 TTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLK 694
             ||  | |||     : |: ||   ::|     |||    ||::     |  |    |
Db     652 KVEEEKTPATPKTEVKESKKDEAPKEAQKP-KAEEKEPLTEKPKDSPGEAKKEEA----K 706

Qy     695 EPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTP 754
           |     |::  | :|      :|     |  |            |||:  : ||     |
Db     707 EKKAAAPEEETPAKLGVKEEAKPKEKAEDAKA-----------KEPSKPSEKE----KP 750

Qy     755 KGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTT 814
           |       ||  |  |:|   ||           ||  |     ||  |    ::
Db     751 K-------KEEVPAAPEKKDTKE---------EKTTESKKREEKPKMEAKAKEED----- 789

Qy     815 PKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDE 851
             |  |  |  |    :| |: |  |:     |:|::
Db     790 --KGLPQEPSKPKTEKAEKSSSTDQKDSQPSEKAPED 824


RESULT  13
MUC1_XENLA
ID   MUC1_XENLA      STANDARD;      PRT;   662 AA.
AC   Q05049;
DT   01-OCT-1994 (Rel. 30, Created)
DT   01-OCT-1994 (Rel. 30, Last sequence update)
DT   01-OCT-1994 (Rel. 30, Last annotation update)
DE   INTEGUMENTARY MUCIN C.1 (FIM-C.1) (FRAGMENT).
OS   Xenopus laevis (African clawed frog).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Amphibia; Batrachia; Anura; Mesobatrachia; Pipoidea; Pipidae;
OC   Xenopodinae; Xenopus.
OX   NCBI_TaxID=8355;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Skin;
RX   MEDLINE=93077556; PubMed=1447205;
RA   Hauser F., Hoffmann W.;
RT   "P-domains as shuffled cysteine-rich modules in integumentary mucin
RT   C.1 (FIM-C.1) from Xenopus laevis. Polydispersity and genetic
RT   polymorphism.";
RL   J. Biol. Chem. 267:24620-24624(1992).
CC   -!- FUNCTION: COULD BE INVOLVED IN DEFENSE AGAINST MICROBIAL
CC       INFECTIONS. PROTECTS THE EPITHELIA FROM EXTERNAL ENVIRONMENT.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- ALTERNATIVE PRODUCTS: A NUMBER OF DIFFERENT FORMS OF THE PROTEIN
CC       MAY BE PRODUCED BY ALTERNATIVE SPLICING.
CC   -!- TISSUE SPECIFICITY: SKIN.
CC   -!- PTM: EXTENSIVELY O-GLYCOSYLATED.
CC   -!- SIMILARITY: CONTAINS 6 P-TYPE (TREFOIL) DOMAINS.
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; L02115; AAA74725.1; -.
DR   PIR; A45155; A45155.
DR   HSSP; P04002; 1WFA.
DR   InterPro; IPR000519; P_trefoil.
DR   Pfam; PF00088; trefoil; 6.
DR   SMART; SM00018; P; 6.
```

```
DR   InterPro; IPR002965; P_rich_extensn.
DR   PRINTS; PR01216; ADHESIVEI.
DR   PRINTS; PR01217; PRICHEXTENSN.
KW   Signal; Repeat; Hydroxylation.
FT   SIGNAL        1     20       POTENTIAL.
FT   CHAIN        21    872       ADHESIVE PLAQUE MATRIX PROTEIN.
FT   DOMAIN       21     41       NONREPETIVE LINKER.
FT   DOMAIN      124    872       TANDEM REPEATS OF Y-K-[PS]-K-[IP]-[ST]-Y-
FT                                P-[PST]-[ST].
FT   DOMAIN      184    192       NONAPEPTIDE 1.
FT   DOMAIN      213    221       NONAPEPTIDE 2.
SQ   SEQUENCE   872 AA;  101677 MW;  98CC70D7C75FF3C4 CRC64;


  Query Match             7.0%;  Score 518.5;  DB 1;  Length 872;
  Best Local Similarity   29.2%;  Pred. No. 6.9e-14;
  Matches  276;  Conservative   80;  Mismatches  407;  Indels  181;  Gaps   54;

Qy    194 VVDEAGSGLDNGDFKVTTPDTST--TQH-------NKVSTSPKITTAKPINPRPSLPPNS 244
          |   | ||   | :| | | :    ::|        ||: | | |  :|   | | |
Db     29 VYGSAYSGASAGAYK-TLPGSHPYGSKHVPVYKPMNKIPT-PYI--SKKSYPAPYKPKGY 84

Qy    245 DTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVL 304
            :|    |   :|           | :|    : |  | : ::    | |       |:
Db     85 YPTKRYQPTYGSKTNYPPIYKPIAKKLSSYKAIKTTYPAYKAKTSYPPSYK-----HKIT 139

Qy    305 AKPTPKAETTTKGPALTTPKEPT-PTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTK 363
            || |      | :| |  ||    |  | | |  ||   | | |     ||   | | |
Db    140 YPPTYK-------PKITYP--PTYKQKPSYPPSYKPKTTYPPTYK-----PKITYPPTYK 185

Qy    364 SAPT-TPKEPAPT--TTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTT--PKK 418
            |: || :|  |   | :|  | |    ||  ::|: |     |  ||   | |  ||
Db    186 RKPSYTPYKPKATYPPTYKPKITYP----PTYKRKPSYT-----PYKPKTTYPPTYKPKI 236

Qy    419 PAPTTPKEPAPTT----PKEPTPTT--PKEPAPTTKEPAPTTPKEPAPT-APKKPAPTTP 471
            |:  |  |        |:  | |  ||   | | :| |: |    ||  ||   | |
Db    237 SYPSIYKPKASYVSSYKSKKTYPPTYKPKISYPPTYKPKPSYP----PTYKPKVTYPPTY 292

Qy    472 KEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPT-----TTKEPAPTTTKSAPTTPK 526
            | | |: |    |  |  |:   ||   ||  |  |:      :|   ||: ||  | |
Db    293 K-PKPSYPPTYKPKITYPPT-YKPKPSYPTPYKQKPSYPPIYKSKSSYPTSYKSKKTYPP 350

Qy    527 EPSPTTTKEPA-PTTPKEPAPTTPKKP-APT-TPK-------EPAPTTPKEPAPTTTKKP 576
            |  |  |     |  |   ||| :|| ||        :| |: |    | || |
Db    351 TYKPKITYPPTYKPKPSYPPSYKPKKTYSPTYKPKITYPPTYKPKPSYPPSYKPKTTYPP 410

Qy    577 APTAPKEPAPTTPKETAP-----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEE 631
            ||   | | |  |        : |   ||   |::     || |: |    |
Db    411 T-YKPKISYPPTYKPKASYVSSYKSKKTYPPTYKPKISYPPTYKPKPSYPPTYKPKITYP 469

Qy    632 PT-PTTPEEPAPTTPKAAAPNTPK-EPAPTTPKEPAPTTP-----KEPAPTT--PKETAP 682
            ||    |  |    ||   | | | :|:  || :  |: |     |   ||:   |:| |
Db    470 PTYKPKPSYPPTYKPKITYPPTYKRKPSYPTPYKQKPSYPPIYKSKSSYPTSYKSKKTYP 529

Qy    683 TT--PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPT-TPKGTAPTTPK 739
            |  || | | | |      ||   |    | ||  ||      :  ||  || : | | |
Db    530 PTYKPKITYPPTYK------PKPSYPPSYKPKTTYPPTYKPKIRYPPTYKPKASYPPTYK 583

Qy    740 EPAPTTPKEPAPTTPKGTAPTTLKEPAPTTP----KKPAPKELAPTTTKGPTSTTSDKPA 795
                  || | | | | | |: |    :||:   :  : :  ||:  | |
Db    584 --------------PKITYPPTYK-PKPSYPTPYKQKPSYPPIYKSKSSYPTAYKSKKTY 628

Qy    796 PTT--PKETAPTTPKEPAPTTPK--KPAPTTPET-------PPPTTSEVSTPTTTKEPTT 844
            | |  || | | | | | |: |   :|  | | |       |    |: | | : :   |
Db    629 PPTYKPKITYPPTYK-PKPSYPPSYRPKITYPPTYKPKKSYPQAYKSKGSYPPSYQPKKT 687

Qy    845 IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKT----TERDL 900
            |    |: :  || |     ||    || |    |||    : | ||    | :
Db    688 Y---PPSYKPKKTYPPTYK-----PKISYPPTYK----TKPSYPASYKRKTSYPPTYKPK 735

Qy    901 RTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPK 960
            : | |  | |     :   | |   |  ::   ::    |    |   :: |     ||
Db    736 ISYPSTYKAKP-----SYPPTYKPKPSYASSYKPKIRYPPTYKPKPSYASSYK-----PK 785

Qy    961 VTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ---KPTKAPKKP-TSTKKPK-T 1015
            :           :   | |||     | |:: ||:    ||  ||    |: |||| |
Db    786 I-------------RYPPTYKPKP----SYASSYKPKIRYPPTYKPKPSYASSYKPKIT 827

Qy   1016 MPRVRKPKTT--PTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQ 1057
               |  ||| :  || :   :  |   |       |     : |:|
Db    828 YPPTYKPKISYPPTYKPKITYPPTYKPKISYPPAYKPKISYPSQ 871


RESULT  12
NFH_RAT
ID   NFH_RAT        STANDARD;      PRT;   831 AA.
AC   P16884; Q63368;
DT   01-AUG-1990 (Rel. 15, Created)
DT   01-FEB-1994 (Rel. 28, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   NEUROFILAMENT TRIPLET H PROTEIN (200 KDA NEUROFILAMENT PROTEIN)
DE   (NEUROFILAMENT HEAVY POLYPEPTIDE) (NF-H) (FRAGMENT).
GN   NEFH OR NFH.
OS   Rattus norvegicus (Rat).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Rattus.
OX   NCBI_TaxID=10116;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Brain;
RX   MEDLINE=89065087; PubMed=3143606;
RA   Breen K.C., Robinson P.A., Wion D., Anderton B.H.;
RT   "Partial sequence of the rat heavy neurofilament polypeptide (NF-H).
RT   Identification of putative phosphorylation sites.";
RL   FEBS Lett. 241:213-218(1988).
RN   [2]
RP   SEQUENCE OF 37-831 FROM N.A.
RX   MEDLINE=88309090; PubMed=2457365;
RA   Dautigny A., Pham-Dinh D., Roussel C., Felix J.M., Nussbaum J.L.,
RA   Jolles P.;
RT   "The large neurofilament subunit (NF-H) of the rat: cDNA cloning and
RT   in situ detection.";
RL   Biochem. Biophys. Res. Commun. 154:1099-1106(1988).
RN   [3]
RP   SEQUENCE OF 1-89 AND 243-313 FROM N.A.
RX   MEDLINE=87080760; PubMed=2878828;
RA   Robinson P.A., Wion D., Anderton B.H.;
RT   "Isolation of a cDNA for the rat heavy neurofilament polypeptide
RT   (NF-H).";
RL   FEBS Lett. 209:203-205(1986).
RN   [4]
RP   SEQUENCE OF 318-831 FROM N.A.
RX   MEDLINE=89184647; PubMed=2928342;
RA   Lieberburg I., Spinner N., Snyder S., Anderson J., Goldgaber D.,
RA   Smulowitz M., Carroll Z., Emanuel B.S., Breitner J., Rubin L.;
RT   "Cloning of a cDNA encoding the rat high molecular weight
RT   neurofilament peptide (NF-H): developmental and tissue expression in
RT   the rat, and mapping of its human homologue to chromosomes 1 and
RT   22.";
RL   Proc. Natl. Acad. Sci. U.S.A. 86:2463-2467(1989).
CC   -!- FUNCTION: NEUROFILAMENTS USUALLY CONTAIN THREE IF PROTEINS: L, M,
CC       AND H WHICH ARE INVOLVED IN THE MAINTENANCE OF NEURONAL CALIBER.
CC       NF-H HAS AN IMPORTANT FUNCTION IN MATURE AXONS THAT IS NOT
CC       SUBSERVED BY THE TWO SMALLER NF PROTEINS.
CC   -!- PTM: THERE ARE A NUMBER OF REPEATS OF THE TRIPEPTIDE K-S-P, NFH IS
CC       PHOSPHORYLATED ON A NUMBER OF THE SERINES IN THIS MOTIF. IT IS
CC       THOUGHT THAT PHOSPHORYLATION OF NFH RESULTS IN THE FORMATION OF
CC       INTERFILAMENT CROSS BRIDGES THAT ARE IMPORTANT IN THE MAINTENANCE
CC       OF AXONAL CALIBER.
CC   -!- PTM: PHOSPHORYLATION SEEMS TO PLAY A MAJOR ROLE IN THE FUNCTIONING
CC       OF THE LARGER NEUROFILAMENT POLYPEPTIDES (NF-M AND NF-H), THE
CC       LEVELS OF PHOSPHORYLATION BEING ALTERED DEVELOPMENTALLY AND
CC       COINCIDENT WITH A CHANGE IN THE NEUROFILAMENT FUNCTION.
CC   -!- SIMILARITY: BELONGS TO THE INTERMEDIATE FILAMENT FAMILY.
CC   -!- CAUTION: REF.1 SEQUENCE DIFFERS FROM THAT SHOWN FROM POSITION 783
CC       ONWARD AND IS LONGER DUE TO A FRAMESHIFT.
```

```
RL   Proc. Natl. Acad. Sci. U.S.A. 90:1536-1540(1993).
CC   -!- FUNCTION: MIGHT FUNCTION AS A CALCIUM-SEQUESTERING "SPONGE" TO
CC       REGULATE THE AMOUNT OF FREE CYTOPLASMIC CALCIUM. IT BINDS 0.3 MOL
CC       OF CA+2 PER MOL OF PROTEIN.
CC   -!- SUBUNIT: HOMODIMER (PROBABLE).
CC   -!- SUBCELLULAR LOCATION: CYTOPLASMIC; HYPODENSE COMPARTMENT.
CC   -!- TISSUE SPECIFICITY: SOMA AND AXONS OF PHOTORECEPTOR CELLS OF
CC       COMPOUND EYES AND OCELLI.
CC   -!- DEVELOPMENTAL STAGE: EXPRESSED EARLY IN PHOTORECEPTOR CELL
CC       DEVELOPMENT.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; L02111; AAA28405.1; -.
DR   EMBL; L05080; AAA28420.1; -.
DR   PIR; A47282; A47282.
DR   FlyBase; FBgn0010218; Cpn.
KW   Calcium-binding.
FT   CONFLICT     36     36       A -> AVAPAVVA (IN REF. 2).
FT   CONFLICT     43     43       I -> T (IN REF. 2).
FT   CONFLICT     64     64       I -> V (IN REF. 2).
FT   CONFLICT     76     76       T -> A (IN REF. 2).
FT   CONFLICT    100    100       P -> PP (IN REF. 2).
FT   CONFLICT    126    127       VQ -> AP (IN REF. 2).
FT   CONFLICT    154    154       I -> V (IN REF. 2).
FT   CONFLICT    160    160       S -> T (IN REF. 2).
FT   CONFLICT    534    534       A -> E (IN REF. 2).
FT   CONFLICT    699    699       I -> T (IN REF. 2).
FT   CONFLICT    703    703       V -> L (IN REF. 2).
FT   CONFLICT    721    721       D -> E (IN REF. 2).
SQ   SEQUENCE   865 AA;  84781 MW;  2110417E0B0E7CFE CRC64;
```

```
  Query Match             7.2%;  Score 530;  DB 1;  Length 865;
  Best Local Similarity  26.2%;  Pred. No. 2.5e-14;
  Matches  233;  Conservative   73;  Mismatches  433;  Indels  152;  Gaps   41;

Qy   298 APTSKVLAKP-TPK---AETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTT 353
         :| |  :| | ||    |      ||  |        |  | : ||  | ||     | |
Db     8 SPVSAPVAAPVTPSAVAAPVQVVSPAAVAP------APAAPIAVTPVAPPPTLASVQPAT 61

Qy   354 PKEPAPTTTKSAPTTP-KEPAPTTTKEPAPTTPKEPAPTTTKEPAPT----TTKSAPTTP 408
            |||    :|  ||    ||     | |     || :    |           |||  |
Db    62 VTIPAPAPIAAASVTPVASVAPPVVAAPTP-----PAASPVSTPVAVAQIPVAVSAPVAP 116

Qy   409 KEPAPTTPKKPAP-------TTPKEPAPTTPKEPTP-------TTPKEPAPTTKEPAPTT 454
            |  ||    |       | |   :  ||   ||       : |  || ||   |
Db   117 PVAATPTPVVQIPVAAPVIATPPVAASAPTPAAVTPVISPVIASPPVVPANTTVPVAAPV 176

Qy   455 PKEPA--PTAPKKPAPTTPKEPAPTTPKEPAPTTTKE-PSPTTPKEPAP-----TTTKSA 506
            ||  |      ||      ||   : |||    | |  | |:  ||      :  :
Db   177 AAVPAAVPVVAPVLAPAVAPAVAPVVAETPAPPPVAEIPVATIPECVAPLIPEVSVVATK 236

Qy   507 PTTTKEP---APTTTKSAPTTPKEPSPTTTKEPA-PTTPKEPAPTTPKKPAPTTPKEPAP 562
         |    ||   ||  |::    |   ||  :  ||  |    |   : : |         ||
Db   237 PLAAAEPVVVAPPATETPVVAPAAASPHVSVAPAVETAVVAPVSASTEPPVAAATLTTAP 296

Qy   563 TTP--------KEPAPTTTKKPAPTAPKEP---APTTPKET--------APTTPKKLTPT 603
           ||          : |  |      ||   ||   ||     ||         | |||  : |
Db   297 ETPALAPVVAESQVAANTVVATPPTPAPEPETIAPPVVAETPEVASVAVAETTPPVVPPV 356

Qy   604 TPEKL-APTTPEKPAPTTPEELAPTTPEEPTPTTPEEP---APTTPKAAAPNTPKEPAPT 659
            | : ||     | |  ||  | |:      || |   ||:   :|   || : ||
Db   357 AAESIPAPVVATTPVPAT---LAVTDPDVTASAVPELPPVIAPSPVPSAVAETPVDLAPP 413

Qy   660 TPKEPAPTTPKEPAPTTPKETAPTTP-KGTAPTTLKE----------PAPTTPKKPAPKE 708
               |    || |    |  | ||   :|| |:            ||:     ||
```

```
Db   414 V----LPPVAAEPVPAVVAEETPETPAPASAPVTIAALDIPEVAPVIAAPSDAPAEAPSA 469

Qy   709 LAPTTTKEPTSTTSDKPAPTTPKGTAPTTP------KEPAPTTPKEP--APTTPKGTAPT 760
            ||  :  |  ||:  |  | |    || |       | |  ||  |    ||    |
Db   470 AAPIVSTPP--TTASVPETTAPPAAVPTEPIDVSVLSEAAIETPVAPPVEVTTEVAVADV 527

Qy   761 TLKEPA------PTTPKKPAPKELAPTTT----KGPTSTTSDKPAPTT----PKETAPTT 806
            | |       |  |  | |  |  ||:    :   ||:|  | | |    | |   :
Db   528 APPEAAADLIIEPVEPPAPIPDLLEQTTSVPAVEAAESTSS--PIPETSLPPPNEAVASP 585

Qy   807 PKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPD----------ESTPEL 856
            || |  :| |      | |: :  : |   : | | :: |                | ||
Db   586 EVAVAPITAPEPIP----EPEPSLATPTEPIPVEAPVVIQEAVDAVEVPVTETSTSIPET 641

Qy   857 SAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTP--ETTTAAPKMT 914
            : |  |:|:     :| :  |: | |:              |     :: ||| | ||| :::
Db   642 TVE-FPEAVAEKVLDPAI--TEAPVTTQEPDVANINDGAPATEI-TTPAVEIVTAAAEVS 697

Qy   915 KETATTTEKTTESKITATTTQVTST------TTQDTTPFKITTLKTTTLA-PKVTTTKKT 967
               :     :|       | |         | | | :    : :  | | ::
Db   698 DIAIPVIDPPVPQEIAVAEIPETDTKPAEVIVEQSTIPIEAPVPEVSKYAEPVISEAPAA 757

Qy   968 ITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPT 1027
               : |: |:          : |  |  : ||   : |  :  |     |   | ||
Db   758 EVPITAGDNPDNTSVGISEVVPTIAEKPVEEVPTS--EIPEQSSSPS--DSVPVAKITPL 813

Qy  1028 PRKMTSTMPELNPTSRIAEAM---LQTTTRPNQTPNSKLVEVN-----PKS 1070
            | : :|   |    :   :|:   |:      ||     :| |:      |||
Db   814 LRDLQTTDVSLLAIAATLDAIGEKLKDQKARNQQVMDRLCEIEKILGPPKS 864
```

```
RESULT  11
FP1_MYTCO
ID   FP1_MYTCO      STANDARD;      PRT;   872 AA.
AC   Q25434;
DT   30-MAY-2000 (Rel. 39, Created)
DT   30-MAY-2000 (Rel. 39, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ADHESIVE PLAQUE MATRIX PROTEIN PRECURSOR (FOOT PROTEIN 1) (MCFP1).
GN   FP1.
OS   Mytilus coruscus (Sea mussel).
OC   Eukaryota; Metazoa; Mollusca; Bivalvia; Pteriomorphia; Mytiloida;
OC   Mytiloidea; Mytilidae; Mytilus.
OX   NCBI_TaxID=42192;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Foot;
RX   MEDLINE=96394686; PubMed=8798340;
RA   Inoue K., Takeuchi Y., Takeyama S., Yamaha E., Yamazaki F., Odo S.,
RA   Harayama S.;
RT   "Adhesive protein cDNA sequence of the mussel Mytilus coruscus and
RT   its evolutionary implications.";
RL   J. Mol. Evol. 43:348-356(1996).
CC   -!- FUNCTION: PROVIDES ADHESIVENESS TO THE MUSSEL'S FOOT. MUSSELS
CC       PRODUCE ONE OF THE STRONGEST WATER INSOLUBLE GLUE. THE MUSSEL'S
CC       ADHESIVE IS A BUNDLE OF THREADS, CALLED A BYSSUS, FORMED BY A
CC       FIBROUS COLLAGENOUS CORE COATED WITH ADHESIVE PROTEINS.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- TISSUE SPECIFICITY: PRODUCED BY THE BYSSAL GLAND.
CC   -!- DOMAIN: ALMOST EXCLUSIVELY COMPOSED OF REPEATS OF A DECAPEPTIDE.
CC   -!- PTM: PROLINES IN THE REPEAT ARE HYDROXYLATED (SINGLE OR DI-) AND
CC       ALSO TYROSINE (THUS PRODUCING DOPA = 3,4-DIHYDROXYPHENYLALANINE).
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; D63777; BAA09850.1; -.
DR   InterPro; IPR002964; Adhesive_plaq.
```

```
DT   01-NOV-1991 (Rel. 20, Created)
DT   01-NOV-1991 (Rel. 20, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   SIALIDASE (EC 3.2.1.18) (NEURAMINIDASE) (NA) (MAJOR SURFACE ANTIGEN).
GN   TCNA.
OS   Trypanosoma cruzi.
OC   Eukaryota; Euglenozoa; Kinetoplastida; Trypanosomatidae; Trypanosoma.
OX   NCBI_TaxID=5693;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=SILVIO X-10/4;
RX   MEDLINE=91277609; PubMed=1711561;
RA   Pereira M.E.A., Mejia J.S., Ortega-Barria E., Matzilevich D.,
RA   Prioli R.P.;
RT   "The Trypanosoma cruzi neuraminidase contains sequences similar to
RT   bacterial neuraminidases, YWTD repeats of the low density lipoprotein
RT   receptor, and type III modules of fibronectin.";
RL   J. Exp. Med. 174:179-191(1991).
RN   [2]
RP   SUBCELLULAR LOCATION.
RX   MEDLINE=91376547; PubMed=1896773;
RA   Prioli R.P., Mejia J.S., Aji T., Aikawa M., Pereira M.E.A.;
RT   "Trypanosoma cruzi: localization of neuraminidase on the surface of
RT   trypomastigotes.";
RL   Trop. Med. Parasitol. 42:146-150(1991).
CC   -!- FUNCTION: DEVELOPMENTALLY REGULATED NEURAMINIDASE IMPLICATED IN
CC       PARASITE INVASION OF CELLS.
CC   -!- CATALYTIC ACTIVITY: HYDROLYSIS OF 2,3-, 2,6- AND 2,8-GLYCOSIDIC
CC       LINKAGES JOINING TERMINAL NON-REDUCING N- OR O-ACYLNEURAMINYL
CC       RESIDUES TO GALACTOSE, N-ACETYLHEXOSAMINE, OR N- OR O-ACYLATED
CC       NEURAMINYL RESIDUES IN OLIGOSACCHARIDES, GLYCOPROTEINS,
CC       GLYCOLIPIDS OR COLOMINIC ACID.
CC   -!- SUBCELLULAR LOCATION: ATTACHED TO THE MEMBRANE BY A GPI-ANCHOR
CC       (POSSIBLE).
CC   -!- DEVELOPMENTAL STAGE: MAXIMAL ACTIVITY IN TRYPOMASTIGOTES, MINIMUM
CC       IN EPIMASTIGOTES AND NO DETECTABLE IN AMASTIGOTES.
CC   -!- MISCELLANEOUS: THE VARIABLE LENGTHS OF THE LONG TANDEM REPEAT
CC       DOMAIN COULD ACCOUNT IN PART FOR THE POLYMORPHISM OF THE TNCA
CC       PROTEIN.
CC   -!- SIMILARITY: BELONGS TO FAMILY 33 OF GLYCOSYL HYDROLASES.
CC   -!- SIMILARITY: CONTAINS 3 BNR REPEATS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M61732; AAA30255.1; -.
DR   PIR; JH0557; JH0557.
DR   HSSP; P29768; 1DIL.
DR   InterPro; IPR002860; BNR.
DR   Pfam; PF02012; BNR; 2.
KW   Hydrolase; Glycosidase; Glycoprotein; Repeat; GPI-anchor;
KW   Phosphorylation.
FT   DOMAIN        1    457       CYS-RICH.
FT   REPEAT       23     34       BNR 1.
FT   REPEAT      163    174       BNR 2.
FT   REPEAT      209    220       BNR 3.
FT   DOMAIN      458    588       FIBRONECTIN TYPE-III.
FT   DOMAIN      589   1120       44 X 12 AA TANDEM REPEATS, LTR DOMAIN.
FT   CARBOHYD    342    342       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    394    394       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1125   1125       N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE   1162 AA;  120032 MW;  07049221897C6A40 CRC64;

  Query Match             7.2%;  Score 530.5;  DB 1;  Length 1162;
  Best Local Similarity   29.4%;  Pred. No. 3e-14;
  Matches  177;  Conservative  100;  Mismatches  212;  Indels  113;  Gaps   34;
```

```
Qy    327 TPTTPKE-PASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTP 385
          ||:|| :  | :||  |  :   || :||  || ::    |:|| : :   |    |:||
Db    599 TPSTPADSSAHSTPSTPVDS---SAHSTPSTPADSSAHGTPSTPVDSSAHGT----PSTP 651

Qy    386 KE-PAPTTTKEPAPTTTKSAPTTPKE-PAPTTPKKP-------APTTPKEPA----PTTP 432
          :  |  |   |  ::  | |:|| :  | :||  |        ||:|| : :    |:||
Db    652 ADSSAHGTPSTPVDSSAHSTPSTPVDSSAHSTPSTPVDSSAHGAPSTPADSSAHGTPSTP 711

Qy    433 KEPTP-TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPS 491
          : :   ||  || ::    |:|| : :        | :||  || ::      |     :
Db    712 VDSSAHGTPSTPADSSAHSTPSTPADSS-------AHSTPSTPADSSAHSTPSTPVDSSA 764

Qy    492 PTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA----PT 547
            ||  || ::  | |:|   || ::    |:|| : |  :|    |:|| : :    |:
Db    765 HGTPSTPADSSAHSTPST---PADSSAHGTPSTPVDSSAHST----PSTPVDSSAHGTPS 817

Qy    548 TP-KKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKL----TP 602
          ||    | :||  |  ::      |     | : |  || ::   | |:||       ||
Db    818 TPVDSSAHSTPSTPVDSSAHGTPSTPVDSSAHSTPSTPADSSAHST-PSTPADSSAHGTP 876

Qy    603 TTP-EKLAPTTPEKPAPTTPEELAPTTPEE----PTPTTPEEPA----PTTP-KAAAPNT 652
          :|| :  | :||  || ::      |:|| :     ||:|| : :    |:||  ::|  |
Db    877 STPVDSSAHSTPSTPADSSAHS-TPSTPVDSSAHSTPSTPADSSAHGTPSTPVDSSAHGT 935

Qy    653 PKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPT 712
          |  || ::    | :||  || ::   | |:||  ::       | :||  |       |
Db    936 PSTPADSS----AHSTPSTPADSSAHST-PSTPADSS-------AHSTPSTPVDSSAHST 983

Qy    713 TTKEPTSTTSDKPAPTTPKGTAPTTPKE-PAPTTPKEPAPTTPKGTAPTTLKEPAPTTPK 771
                || :|  | :|     |:|| :  | :||  || ::    |  | :  | :||
Db    984 -----PSTPADSSAHST-----PSTPADSSAHSTPSTPADSSAHSTPSTPVDSSAHSTPS 1033

Qy    772 KPAPKELAPTTTKGPTSTTSDKPAPTTPK----ETAPTTPKEPAPTTPKKPAPTTPETPP 827
          |      | ::  |  || :|  | :||       :| :||  || ::    |  || ||
Db   1034 TP-----ADSSAHGTPSTPADSSAHSTPSTPVDSSAHSTPSTPADSS----AHGTPSTPA 1084

Qy    828 PTTSEVSTPTTTKEPTTIHKSPDESTP-ELSAEPTPKALENSPKEPGVPTTKTPAATKPE 886
          :::  |||:|  : :: | :|  ||| : ||  ||                 |||  :
Db   1085 DSSAH-STPSTPAD-SSAHGTP--STPADSSAHSTP--------------STPAGSSAN 1125

Qy    887 MT 888
          |
Db   1126 GT 1127
```

```
RESULT  10
CPN_DROME
ID   CPN_DROME      STANDARD;      PRT;   865 AA.
AC   Q02910;
DT   01-OCT-1993 (Rel. 27, Created)
DT   01-OCT-1993 (Rel. 27, Last sequence update)
DT   01-FEB-1994 (Rel. 28, Last annotation update)
DE   CALPHOTIN.
GN   CPN OR CAP.
OS   Drosophila melanogaster (Fruit fly).
OC   Eukaryota; Metazoa; Arthropoda; Tracheata; Hexapoda; Insecta;
OC   Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha;
OC   Ephydroidea; Drosophilidae; Drosophila.
OX   NCBI_TaxID=7227;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=CANTON-S;
RX   MEDLINE=93165729; PubMed=8094559;
RA   Martin J.H., Benzer S., Rudnicka M., Miller C.A.;
RT   "Calphotin: a Drosophila photoreceptor cell calcium-binding protein.";
RL   Proc. Natl. Acad. Sci. U.S.A. 90:1531-1535(1993).
RN   [2]
RP   SEQUENCE FROM N.A.
RC   STRAIN=CANTON-S;
RX   MEDLINE=93165730; PubMed=8434015;
RA   Ballinger D.G., Xue N., Harshman K.D.;
RT   "A Drosophila photoreceptor cell-specific protein, calphotin, binds
RT   calcium and contains a leucine zipper.";
```

```
RL   Gene 68:307-314(1988).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=89089138; PubMed=3145094;
RA   Shneidman P.S., Carden M.J., Lees J.F., Lazzarini R.A.;
RT   "The structure of the largest murine neurofilament protein (NF-H) as
RT   revealed by cDNA and genomic sequences.";
RL   Brain Res. 464:217-231(1988).
RN   [3]
RP   SEQUENCE FROM N.A.
RC   STRAIN=SWISS WEBSTER; TISSUE=Brain;
RA   Carden M.J.;
RL   Submitted (MAR-1994) to the EMBL/GenBank/DDBJ databases.
CC   -!- FUNCTION: NEUROFILAMENTS USUALLY CONTAIN THREE IF PROTEINS: L, M,
CC       AND H WHICH ARE INVOLVED IN THE MAINTENANCE OF NEURONAL CALIBER.
CC       NF-H HAS AN IMPORTANT FUNCTION IN MATURE AXONS THAT IS NOT
CC       SUBSERVED BY THE TWO SMALLER NF PROTEINS.
CC   -!- PTM: THERE ARE A NUMBER OF REPEATS OF THE TRIPEPTIDE K-S-P, NFH IS
CC       PHOSPHORYLATED ON A NUMBER OF THE SERINES IN THIS MOTIF. IT IS
CC       THOUGHT THAT PHOSPHORYLATION OF NFH RESULTS IN THE FORMATION OF
CC       INTERFILAMENT CROSS BRIDGES THAT ARE IMPORTANT IN THE MAINTENANCE
CC       OF AXONAL CALIBER.
CC   -!- PTM: PHOSPHORYLATION SEEMS TO PLAY A MAJOR ROLE IN THE FUNCTIONING
CC       OF THE LARGER NEUROFILAMENT POLYPEPTIDES (NF-M AND NF-H), THE
CC       LEVELS OF PHOSPHORYLATION BEING ALTERED DEVELOPMENTALLY AND
CC       COINCIDENT WITH A CHANGE IN THE NEUROFILAMENT FUNCTION.
CC   -!- SIMILARITY: BELONGS TO THE INTERMEDIATE FILAMENT FAMILY.
CC   -!- CAUTION: REF.2 SEQUENCE DIFFERS FROM THAT SHOWN IN POSITIONS 534
CC       TO 716 AND IS SHORTER DUE TO FRAMESHIFTS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M24496; AAA39813.1; -.
DR   EMBL; M23349; AAA39813.1; JOINED.
DR   EMBL; M24494; AAA39813.1; JOINED.
DR   EMBL; M24495; AAA39813.1; JOINED.
DR   EMBL; M35131; AAA39809.1; ALT_FRAME.
DR   EMBL; Z31012; CAA83229.1; -.
DR   PIR; JT0368; QFMSH.
DR   PIR; A43778; A43778.
DR   MGD; MGI:97309; Nfh.
DR   InterPro; IPR001664; IF.
DR   Pfam; PF00038; filament; 1.
DR   PROSITE; PS00226; IF; 1.
KW   Intermediate filament; Coiled coil; Neurone; Phosphorylation;
KW   Repeat.
FT   DOMAIN        1     97       HEAD.
FT   DOMAIN       98    408       ROD.
FT   DOMAIN      409   1087       TAIL.
FT   DOMAIN      436    517       GLU-RICH (ACIDIC).
FT   DOMAIN      519    886       50 X 6 AA TANDEM REPEATS OF K-S-P-A-E-A.
FT   DOMAIN      887   1087       GLU/LYS-RICH.
FT   DOMAIN       98    129       COIL 1A.
FT   DOMAIN      130    141       LINKER 1.
FT   DOMAIN      142    239       COIL 1B.
FT   DOMAIN      240    261       LINKER 12.
FT   DOMAIN      262    283       COIL 2A.
FT   DOMAIN      284    287       LINKER 2.
FT   DOMAIN      288    408       COIL 2B.
FT   CONFLICT    133    133       K -> QA (IN REF. 2 AND 3).
FT   CONFLICT    199    199       A -> AR (IN REF. 2 AND 3).
FT   CONFLICT    281    281       S -> T (IN REF. 2 AND 3).
FT   CONFLICT    492    492       L -> G (IN REF. 2 AND 3).
FT   CONFLICT    551    551       P -> PREAKSP (IN REF. 3).
FT   CONFLICT    689    712       MISSING (IN REF. 3).
FT   CONFLICT    714    714       G -> A (IN REF. 3).
FT   CONFLICT    814    814       V -> M (IN REF. 2 AND 3).
FT   CONFLICT    843    843       T -> N (IN REF. 2 AND 3).
SQ   SEQUENCE   1087 AA;  116612 MW;  57BAC76A38ED1CB9 CRC64;


  Query Match             7.4%;  Score 551;  DB 1;  Length 1087;
  Best Local Similarity   26.8%;  Pred. No. 4.6e-15;
  Matches  234;  Conservative   96;  Mismatches  358;  Indels  186;  Gaps   43;

Qy     119 KKTKKVIESE--EITEEHSVS-ENQESSSSSSSSSSSSTIWKIKSSKNSAANRELQKKLK 175
           | ||: :| :  |: : |     : : :     |   :| |::     :|  || |  |
Db     331 KSTKESLERQRSELEDRHQADIASYQDAIQQLDSELRNTKWEM-----AAQLREYQDLLN 385

Qy     176 VK--------------DNKKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNK 221
           ||              : :: |    |:|          |  |  |:  |  ||  | |
Db     386 VKMALDIEIAAYRKLLEGEECRIGFGPSP---------FSLTEGLPKI--PSIST--HIK 432

Qy     222 VSTSPKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTT 281
           | :  |   :             : ||| :   :   :   |  | :     |:|
Db     433 VKSEEMIKVVE-----------KSEKETVIVEGQTEEIRVTEGVTEEEDKEAQGQEGEE 480

Qy     282 SAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKE 341
           : :  :  |:  |   :| ::  | |  : ::  |  | :  :  :|   | ||    :
Db     481 AEEGEEKEEEELAAATSPPAEEAASPEKETKSRVKEEAKSPGEAKSPGEAKSPA----EA 536

Qy     342 PTPTTIKSAPTTPKEPAPTTTKSAPTTPKEP-APTTTKEPAPTTPKEPAPTTTKEP---- 396
           :|   || |  | |    : : | :| || :|   | ||   || ||  | | |
Db     537 KSPGEAKS-PGEAKSPGEAKSPAEPKSPAEPKSPAEAKSPA--EPKSPA--TVKSPGEAK 591

Qy     397 APTTTKSAPTTPKEPA----PTTPKKPA----PTTPKEPA----PTTPKEP----TPTTP 440
           :|:  || |   | ||    |   | ||    |   | ||    | | | |    :|:
Db     592 SPSEAKS-PAEAKSPAEAKSPAEAKSPAEAKSPAEAKSPATVKSPGEAKSPSEA 650

Qy     441 KEPAPTTKEPAPTTPKEPA----PTAPKKP----APTTPKEPAPTTPKEPAPTTTKEPSP 492
           | ||   :  :|  | .||    |   | |    :|  || ||    | ||    :  ||
Db     651 KSPA---EAKSPAEAKSPAEAKSPAEVKSPGEAKSPAEPKSPAEA--KSPA----EVKSP 701

Qy     493 TTPKEPA----PTTTKSAPTTTKEPAPTTTKSAPTTPKE-PSPTTTKEPA----PTTPKE 543
             | ||    |   || |   | ||   : :|  :| |  ||   | ||    |   |
Db     702 AEAKSPAEVKSPGEAKS-PAAVKSPAEAKSPAAVKSPGEAKSPGEAKSPAEAKSPAEAKS 760

Qy     544 PAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPT 603
           |       : | |  || | :  :  :|   | |    | ||     | |  :| | |
Db     761 PIEVKSPEKAKTPVKEGAKSPAEAKSPEKAKSPVKEDIKPPAEAKSPEKA-KSPVKEGAK 819

Qy     604 TPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEE-PAPTTPKAAAPNTPKEPAPTT-- 660
           |||  |   : |   || :   | | :  |  ||:  :|   || :|   || | |:
Db     820 PPEKAKPLDVKSPEAQTPVQEEATVPTDIRP--PEQVKSPAKEKAKSPE--KEEAKTSEK 875

Qy     661 --PKEPAPTTP-------KEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAP 711
             ||:    :|       |||     :|    |||  |  : |: ||    ||  :|
Db     876 VAPKKEEVKSPVKEEVKAKEPPKKVEEEKTLPTPKTEAKESKKDEAPKEAPKPKVEEKKE 935

Qy     712 TTTKEPTSTTSD--KPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTT 769
           | |::|  :|::  |       |    :  : ||    || |   ||    ||  |   |
Db     936 TPTEKPKDSTAEAKKEEAGEKKKAVASEEETPAKLGVKEEA--KPKEKTETTKTEAEDTK 993

Qy     770 PKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETPPPT 829
           |:|:    || |: |                    ||  |  |:|
Db     994 AKEPS----KPTETEKP-------------------KKEEMPAAPEK------------ 1017

Qy     830 TSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTT 889
                 | :| ||  : |:|           || :|   ||     :| |  :||:
Db    1018 ------KDTKEEKTTESRKPEEK---------PK-MEAKVKEDDKSLSKEP--SKPKTEK 1059

Qy     890 TAKDKTTERDLRTTPETTTAAPKMTKETATTTEK 923
             |  :|::     ||      | |:: ||  ||
Db    1060 AEKSSSTDQKESQPPE------KTTEDKATKGEK 1087


RESULT   9
TCNA_TRYCR
ID   TCNA_TRYCR     STANDARD;      PRT;  1162 AA.
AC   P23253;
```

```
Qy     582 KEPAPTTPKETAPT-----TPKKLTPTTPEKLA-----PTTPEKPAPTTPEELAPTTPEE 631
           ::|  :| | | ||     :|:|||  | |||      || | :    : |:|  | | |
Db     758 EKPTISTEKPTIPTEKPTISPEKLTIPT-EKLTIPTEKPTIPIEETTISTEKL--TIPTE 814

Qy     632 PTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPT 691
             :||:|  :| |           || | |  || | |       ||  :| | | ||
Db     815 KPTISPEKPTISTEK----------PTIPTE-KPTIPTE-------ETTISTEKLTIPT 855

Qy     692 TLKEPAPTTPKKPAPKELAPTTTKEPTSTT------SDKPAPTTPKGTAPTTPKEPAPTT 745
             ::|  :  |   | |    :|:::||  |       ::||   | | | ||  ::
Db     856 --EKPTISPEKLTIPTEKPTISTEKPTIPTEKLTIPTEKPTIPTEKPTIPT--EKLTALR 911

Qy     746 PKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKET--A 803
           |  |:||   | |   :   ||:||       :  |||   :||    | |    |:
Db     912 PPHPSPTA-TGLAALVMSPHAPSTPMTSV---ILGTTTTSRSSTGMSCP-PNARYESCAC 966

Qy     804 PTTPKEPAPT 813
           | : | | |:
Db     967 PASCKSPRPS 976


RESULT   7
EXTN_TOBAC
ID   EXTN_TOBAC     STANDARD;      PRT;   620 AA.
AC   P13983;
DT   01-JAN-1990 (Rel. 13, Created)
DT   01-JAN-1990 (Rel. 13, Last sequence update)
DT   01-MAR-1992 (Rel. 21, Last annotation update)
DE   EXTENSIN PRECURSOR (CELL WALL HYDROXYPROLINE-RICH GLYCOPROTEIN).
GN   HRGPNT3.
OS   Nicotiana tabacum (Common tobacco).
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC   Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots;
OC   Asteridae; euasterids I; Solanales; Solanaceae; Nicotiana.
OX   NCBI_TaxID=4097;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=CV. XANTHI; TISSUE=Leaf;
RX   MEDLINE=90128263; PubMed=2612909;
RA   Keller B., Lamb C.J.;
RT   "Specific expression of a novel cell wall hydroxyproline-rich
RT   glycoprotein gene in lateral root initiation.";
RL   Genes Dev. 3:1639-1646(1989).
CC   -!- FUNCTION: HAS A SPECIALIZED STRUCTURAL FUNCTION, POSSIBLY IN
CC       THE MECHANICAL PENETRATION OF THE CORTEX AND EPIDERMIS OF THE
CC       MAIN ROOT.
CC   -!- SUBCELLULAR LOCATION: EXTRACELLULAR MATRIX.
CC   -!- PTM: EXTENSINS CONTAIN A CHARACTERISTIC REPEAT OF THE PENTAPEPTIDE
CC       SER-PRO(4). THE PROLINE RESIDUE IS HYDROXYLATED AND THEN
CC       GLYCOSYLATED.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X13885; CAA32090.1; -.
DR   PIR; S06733; S06733.
KW   Repeat; Cell wall; Glycoprotein; Signal; Structural protein;
KW   Hydroxylation.
FT   SIGNAL       1     ?
FT   CHAIN        ?    620       EXTENSIN.
FT   REPEAT      70     73       H-A-P-P.
FT   REPEAT     148    151       H-A-P-P.
FT   DOMAIN     229    242       2 X 7 AA TANDEM REPEATS OF T-H-R-H-A-P-P.
FT   REPEAT     229    235       1.
FT   REPEAT     236    242       2.
FT   DOMAIN     205    620       CONTAINS THE SER-PRO(4) REPEATS.
FT   DOMAIN     499    600       3 X APPROXIMATE TANDEM REPEATS.

SQ   SEQUENCE   620 AA;  65406 MW;  641DD2278AB28524 CRC64;


  Query Match             7.4%;  Score 551;  DB 1;  Length 620;
  Best Local Similarity   27.2%;  Pred. No. 2.9e-15;
  Matches  170;  Conservative   64;  Mismatches  304;  Indels   88;  Gaps   24;

Qy     311 AETTTK-GPALTTP--KEPTPTT-----PKEPASTTPKE---PTPTTIKSAPTTPKEPAP 359
           || ||: |  |  |   :| |::     |  | :| |     |:|   : ||     | |
Db      24 AEATTQYGGYLPPPVTSQPPPSSIGLSPPSAPTTTPPSRGHVPSP---RHAPPRHAYPPP 80

Qy     360 TTTKSAPTTPKEP------APTTTKEPAPT---TPKEPAPTTTKEPAPTTTKSAPTTPKE 410
           :     |:    |      |:    | |:   :|  | |:    |        |:  :
Db      81 SHGHLPPSVGGPPPHRGHLPPSRGFNPPPSPVISPSHPPPSYGAPPPSHGPGHLPSHGQR 140

Qy     411 PAPTTPKKPAPT---TPKEPAPTTPKEPTPTTPKEPAPTTK--EPAPTTPKEPAPTAPKK 465
           |       |:|:    |     | |:   | : : |:| ::   | | |  :| ||
Db     141 P-------PSPSHGHAPPSGGHTPPRGQHPPSHRRPSPPSRHGHPPPPTYAQPPPTPIYS 193

Qy     466 PAPTTPKEPAPTTPKEPAPTTTKEPSPTTPK---EPAPTTTKSAPTTTKEPAPTTTKS-- 520
           |:|    :| ||    | | |  :|:|: |    :| | | : || | :   ||   |
Db     194 PSPQV--QPPPT--YSPPPPTHVQPTPSPPSRGHQPQPPTHRHAPPTHRHAPPTHQPSPL 249

Qy     521 ---APTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPT---TPKEPAPTTTK 574
             |:  ::| | |    | |   : || |:    | | |   | |:   :|  || : :
Db     250 RHLPPSPRRQPQPPTYSPPPPAYAQSPQPSPTYSPPPPTYSPPPPSPIYSPPPPAYSPSP 309

Qy     575 KPAPT---APKEPAPTTPKETAPTTPKKL-TPTT------PEKLAPTTPEKPAPTTPEEL 624
           | ||   :|  || : |   :|  |  |  |::      |   :|  |   :|  |  |
Db     310 PPTPTPTFSPPPPAYSPPPTYSPPPPTYLPLPSSPIYSPPPPVYSPPPPPSYSPPPPTYL 369

Qy     625 APTTPEEPTPTTPEEPAPTTPKAAAP---NTPKEPAPTTPKEPAPTTPKEPAPTTPKETA 681
           |  |  | | :   | ||  ::  |    :|  |||  |   | ||    | | |  :
Db     370 PPPPPSSPPPPSFSPPPPTYEQSPPPPPAYSPPLPAPPTYSPPPPT--YSPPPPTYAQPP 427

Qy     682 PTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEP 741
           |       | |   | |     | |    | |    | |   |      :| | |  | : |
Db     428 P------LPPTYSPPPPAYSPPPPPTYSPPPPTYSPPPPAYAQPPPPPPTYSPPPPAYSP 481

Qy     742 APTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKE 801
           | :|    |       | |   | |     | |     |   : || |    |:| |
Db     482 PPPSPIYSPPPPQVQPLPPTFSPPPPRRIHLPPPPHRQP---RPPTPTYGQPPSPPTFSP 538

Qy     802 TAPTTPKEPAPT--TPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAE 859
             |     | |    |: | ||  : | | |     |     |   |: |   ||
Db     539 PPPRQIHSPPPPHWQPRTPTPTYGQPPSPPTFSAPPPRQIHSPPPPHRQPRPPTPTYGQP 598

Qy     860 PTPKALENSPKEPGVPTTKTPAATKP 885
           |:|           ||| :| :  |
Db     599 PSP-----------PTTYSPPSPPP 612


RESULT   8
NFH_MOUSE
ID   NFH_MOUSE      STANDARD;      PRT;  1087 AA.
AC   P19246; Q61959;
DT   01-NOV-1990 (Rel. 16, Created)
DT   01-AUG-1992 (Rel. 23, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   NEUROFILAMENT TRIPLET H PROTEIN (200 KDA NEUROFILAMENT PROTEIN)
DE   (NEUROFILAMENT HEAVY POLYPEPTIDE) (NF-H).
GN   NEFH OR NFH.
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=89121513; PubMed=3220257;
RA   Julien J.-P., Cote F., Beaudet L., Sidky M., Flavell D., Grosveld F.,
RA   Mushynski W.;
RT   "Sequence and structure of the mouse gene coding for the largest
RT   neurofilament subunit.";
```

```
Qy     913 MTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTE 972
           |  :    |   | ::|        :   |     |    | |: :: |     |||   |
Db     748 -TYKAKPTYPPTYKAK-------PSYPPTYKPKPSYPPTYKSKSIYPSSYKPKKTYPPT- 798

Qy     973 IMNKPEETAKPKDRATNSKATTPKPQKPTKAPKK---PTSTKKPKTMPRVRKPKTTPTP- 1028
             ||: |  |  :        |||  |      |   |:: |   : |   | ||:  |
Db     799 --YKPKLTYPPTYK--------PKPSYPPSYKPKITYPSTYKLKPSYPPTYKSKTSYPPT 848

Qy    1029 -----------RKMTSTMPELNPTSR 1043
                      :  ||  |   ||:|
Db     849 YNKKISYPSSYKAKTSYPPAYKPTNR 874


RESULT   6
ZAN_HUMAN
ID   ZAN_HUMAN      STANDARD;      PRT;  2700 AA.
AC   Q9Y493; O00218;
DT   20-AUG-2001 (Rel. 40, Created)
DT   20-AUG-2001 (Rel. 40, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ZONADHESIN (FRAGMENT).
GN   ZAN.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE OF 1-2379 FROM N.A.
RX   MEDLINE=99018118; PubMed=9799793;
RA   Glockner G., Scherer S., Schattevoy R., Boright A., Weber J.,
RA   Tsui L.C., Rosenthal A.;
RT   "Large-scale sequencing of two regions in human chromosome 7q22:
RT   analysis of 650 kb of genomic sequence around the EPO and CUTL1 loci
RT   reveals 17 genes.";
RL   Genome Res. 8:1060-1073(1998).
RN   [2]
RP   SEQUENCE OF 2338-2700 FROM N.A.
RC   TISSUE=Testis;
RX   MEDLINE=97271566; PubMed=9126492;
RA   Gao Z., Harumi T., Garbers D.L.;
RT   "Chromosome localization of the mouse zonadhesin gene and the human
RT   zonadhesin gene (ZAN).";
RL   Genomics 41:119-122(1997).
CC   -!- FUNCTION: BINDS IN A SPECIES-SPECIFIC MANNER TO THE ZONA PELLUCIDA
CC       OF THE EGG. MAY BE INVOLVED IN GAMETE RECOGNITION AND/OR
CC       SIGNALING.
CC   -!- SUBUNIT: PROBABLY FORMS COVALENT OLIGOMERS.
CC   -!- SUBCELLULAR LOCATION: TYPE I MEMBRANE PROTEIN, EXCLUSIVELY ON THE
CC       APICAL REGION OF THE SPERM HEAD (BY SIMILARITY).
CC   -!- TISSUE SPECIFICITY: IN TESTIS; PRIMARILY IN HAPLOID SPERMATIDS.
CC   -!- DOMAIN: THE MAM DOMAINS PROBABLY MEDIATE SPERM ADHESION TO THE
CC       ZONA PELLUCIDA.
CC   -!- DOMAIN: DURING SPERM MIGRATION THROUGH THE REPRODUCTIVE TRACTS,
CC       THE MUCIN-LIKE DOMAIN MIGHT INHIBIT INAPPROPRIATE TRAPPING OF
CC       SPERMATOZOA OR PROMOTING ADHESION TO THE OVIDUCTAL ISTHMUS.
CC   -!- DOMAIN: THE VWFD DOMAINS 2 AND 3 MAY MEDIATE COVALENT
CC       OLIGOMERIZATION (BY SIMILARITY TO HUMAN INTESTINAL MUCIN MUC2).
CC   -!- SIMILARITY: CONTAINS 3.5 MAM DOMAINS.
CC   -!- SIMILARITY: CONTAINS 4.5 VWFD DOMAINS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; AF053356; AAC78790.1; -.
DR   EMBL; U83191; AAC51208.1; -.
DR   MIM; 602372; -.
DR   InterPro; IPR000561; EGF-like.
DR   InterPro; IPR000998; MAM.
DR   InterPro; IPR002965; P_rich_extensn.
DR   InterPro; IPR002919; TIL.
DR   InterPro; IPR003328; TILa.
DR   InterPro; IPR001007; VWFC.
DR   InterPro; IPR001846; Vwd.
DR   Pfam; PF00629; MAM; 3.
DR   Pfam; PF01826; TIL; 4.
DR   Pfam; PF02345; TILa; 4.
DR   Pfam; PF00094; vwd; 4.
DR   PRINTS; PR01217; PRICHEXTENSN.
DR   SMART; SM00137; MAM; 2.
DR   SMART; SM00214; VWC; 1.
DR   SMART; SM00011; VWC_def; 3.
DR   SMART; SM00216; VWD; 3.
DR   PROSITE; PS01186; EGF_2; 3.
DR   PROSITE; PS00740; MAM_1; 1.
DR   PROSITE; PS50060; MAM_2; 4.
KW   Glycoprotein; Transmembrane; Cell adhesion; Repeat.
FT   NON_TER       1      1
FT   DOMAIN       <1    109       MAM 1.
FT   DOMAIN      112    136       MAM 2 (PARTIAL).
FT   DOMAIN      161    326       MAM 3.
FT   DOMAIN      322    446       MAM 4.
FT   DOMAIN      483    951       66 X HEPTAPEPTIDE REPEATS (APPROXIMATE)
FT                                (MUCIN-LIKE DOMAIN).
FT   DOMAIN      953   1065       VWFD 1 (PARTIAL).
FT   DOMAIN     1066   1454       VWFD 2.
FT   DOMAIN     1455   1861       VWFD 3.
FT   DOMAIN     1862   2292       VWFD 4.
FT   DOMAIN     2293   2684       VWFD 5.
FT   DOMAIN        ?      ?       EGF-LIKE.
FT   CARBOHYD     74     74       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    403    403       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1023   1023       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1099   1099       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1618   1618       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1737   1737       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1832   1832       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1878   1878       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   2136   2136       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   2505   2505       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CONFLICT   2374   2379       NNQKMA -> RAGPGP (IN REF. 1).
FT   NON_TER    2700   2700
SQ   SEQUENCE   2700 AA;  293013 MW;  80E60CC0B12277B1 CRC64;


  Query Match             7.5%;  Score 555.5;  DB 1;  Length 2700;
  Best Local Similarity   32.6%;  Pred. No. 6.4e-15;
  Matches  199;  Conservative   64;  Mismatches  240;  Indels  107;  Gaps   32;

Qy     233 PINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKT 292
           |:  | ||| |  |     |   |||  |:  | : |             | | ||||
Db     445 PVKVLPELPPVSPVSS----TGPSETTGLTENPTISTK-----------KPTVSIEKP 487

Qy     293 SAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPT-----PTTI 347
           |     ||      |  |     |: | ::| |   || | |  :   ::||     || :
Db     488 SVTTEKPT-----VPKEKPTIPTEKPTISTEK---PTIPSEKPNMPSEKPTIPSEKPTIL 539

Qy     348 KSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPA-PTTTKSAPT 406
             || | |  ||     ||   |     |:|  |||| |   |  :||  ||   | ||
Db     540 TEKPTIPSE-KPTIPSEKPTISTEKPTVPTEE--PTTPTEETTTYMEEPVIPTEKPSIPT 596

Qy     407 TPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKP 466
             ::|:  | |   ||   |    : ::||     |   || |   || | | :  :|:|
Db     597 --EKPSIPTEK---PTISMEETIISTEKPTICPEKPTIPTEK---PTIPTEKSTISPEK- 647

Qy     467 APTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPK 526
           |||| |  || | |    :|::|:  | | |  :  |    | |   ||     || |
Db     648 -PTTPTE-KPTIPTEKPTISTEKPTIPTEK-PTISPEKLTIPTEKLTIPT---EKPTIPT 701

Qy     527 EPSPTTTKEPAPTTPKEPAPTTPKKPA-----PTTPKEPAPTTPKEPAPTTTKKPAPTAP 581
           |    :|:|  |||| |    : :||:     || | |   |: :|   :| |   |
Db     702 EKPTISTEE--PTTPTEETTISTEKPSIPMEKPTLPTEETTTSVEETTISTEKLTIPM-- 757
```

```
Qy     888 TTTAKDKTTERDLRTTPETT-------TAAPKMTKETATTTEKTTESKITATTTQVTSTT 940
Db     754 --PAHGVTSAPDTRPAPGSTAPPAHGVTSAP----DTRPAPGSTAPPAHGVTSAPDTRPA 807

Qy     941 TQDTTP--FKITTLKTTTLAPKVTTTKKTITTT--EIMNKPEETAKPKDRATNSKATTPK 996
Db     808 PGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPA 867

Qy     997 PQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPN 1056
Db     868 P--GSTAPPAHGVTSAPDTRP---APGSTAPPAHGVTSAPDTRPA------------PG 909

Qy    1057 QT-PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPML 1115
Db     910 STAPPAHGVTSAPDTRPAPGSTAPPAH----------GVTSAPDNRPALGSTA----PPV 955

Qy    1116 SDETNICNGKPVDGLTTLRNGTLVAFRGHYFWMLSPFSPPS 1156
Db     956 HNVTSASGSASGSASTLVHNGTSARATTTPASKSTPFSIPS 996
```

```
RESULT   5
FP1_MYTED
ID   FP1_MYTED      STANDARD;      PRT;   875 AA.
AC   Q25460;
DT   30-MAY-2000 (Rel. 39, Created)
DT   30-MAY-2000 (Rel. 39, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ADHESIVE PLAQUE MATRIX PROTEIN (POLYPHENOLIC ADHESIVE PROTEIN) (FOOT
DE   PROTEIN 1) (MEFP1) (FRAGMENT).
GN   FP1.
OS   Mytilus edulis (Blue mussel).
OC   Eukaryota; Metazoa; Mollusca; Bivalvia; Pteriomorphia; Mytiloida;
OC   Mytiloidea; Mytilidae; Mytilus.
OX   NCBI_TaxID=6550;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=91025829; PubMed=1367451;
RA   Filpula D.R., Lee S.M., Link R.P., Strausberg S.L., Strausberg R.L.;
RT   "Structural and functional repetition in a marine mussel adhesive
RT   protein.";
RL   Biotechnol. Prog. 6:171-177(1990).
RN   [2]
RP   PARTIAL SEQUENCE, AND POST-TRANSLATIONAL MODIFICATIONS.
RX   MEDLINE=83135732; PubMed=6298211;
RA   Waite J.H.;
RT   "Evidence for a repeating 3,4-dihydroxyphenylalanine- and
RT   hydroxyproline-containing decapeptide in the adhesive protein of the
RT   mussel, Mytilus edulis L.";
RL   J. Biol. Chem. 258:2911-2915(1983).
CC   -!- FUNCTION: PROVIDES ADHESIVENESS TO THE MUSSEL'S FOOT. MUSSELS
CC       PRODUCE ONE OF THE STRONGEST WATER INSOLUBLE GLUE. THE MUSSEL'S
CC       ADHESIVE IS A BUNDLE OF THREADS, CALLED A BYSSUS, FORMED BY A
CC       FIBROUS COLLAGENOUS CORE COATED WITH ADHESIVE PROTEINS.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- TISSUE SPECIFICITY: PRODUCED BY THE BYSSAL GLAND.
CC   -!- DOMAIN: ALMOST EXCLUSIVELY COMPOSED OF REPEATS OF A DECAPEPTIDE.
CC   -!- PTM: THE DECAPEPTIDE A-K-P-S-Y-P-P-T-Y-K IS POST-TRANSLATIONALLY
CC       MODIFIED AS FOLLOWING: THE SIXTH AND SEVENTH RESIDUES ARE
CC       HYDROXYLATED AND THE PENULTIMATE IS A 3,4-DIHYDROXYPHENYLALANINE
CC       (DOPA) DERIVED FROM TYROSINE.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X54422; CAA38294.1; -.
DR   InterPro; IPR002964; Adhesive_plaq.
DR   InterPro; IPR002965; P_rich_extensn.
DR   PRINTS; PR01216; ADHESIVEI.
DR   PRINTS; PR01217; PRICHEXTENSN.
KW   Repeat; Hydroxylation.
FT   NON_TER       1      1
FT   DOMAIN       67    870       TANDEM REPEATS OF Y-K-[PA]-K-[LP]-[ST]-Y-
FT                                 P-P-[ST].
SQ   SEQUENCE   875 AA;  100412 MW;  6EA85312748CAACE CRC64;
```

```
Query Match             7.5%;  Score 559;  DB 1;  Length 875;
Best Local Similarity   26.9%;  Pred. No. 1.9e-15;
Matches  281;  Conservative  102;  Mismatches  403;  Indels  260;  Gaps   56;

Qy      84 HNPTSPPSSKKAPPPSGASQTIKSTTKRSPKPPNKKKTKKVIESEEITEEHSVSENQESS 143
Db       3 HEPVYKP-------------VKTSYSAPYKPPTYQPLKKKVD-------------YRPT 35

Qy     144 SSSSSSSSSSTIWKIKSSKNSAANRELQKKLKVKDNKKNRTKKKPTPKPPVVDEAGSGLD 203
Db      36 KSYPPTYGSKTNYLPLAKKLSS-----YKPIKTTYNAK-------TNYPPVY-------- 75

Qy     204 NGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETK 263
Db      76 ---------------------KPKMTYPPTYKPKPSYPP----------TYKSKPTYKPK 104

Qy     264 ETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTP 323
Db     105 IT-----------YPPTYKAKPSYPSSYKPKKTYPPTYK------PKLTY----PPTYKP 143

Qy     324 KEPTPTT----PKEPASTTPKEPTPTTIKSAPTTP-----KEPAPTTTKSAPTTPKEPAP 374
Db     144 KPSYPPTYKPKPSYPPSYKTKKTYPSSYKAKPSYPPTYKAKPSYPPTYKAKPSYP----P 199

Qy     375 TTTKEPA-PTTPKEPAPTTTKEPAPTTTKSAPTTPKEPA--PTTPKKPA--PTTPKEPA- 428
Db     200 TYKAKPTYKAKPTYPSTYKAKPSYPPTYKAKPTYKAKPSYPPTYKAKPSYPPTYKAKPSY 259

Qy     429 -PTTPKEPT-PTTPKEPAPTTKEPAPTTPKEPAPTAPKKPA--PTTPKEPA--PTTPKEP 482
Db     260 PPTYKAKPTYKAKPTYKAKPTYKAKPSYP----PTYKAKPSYPPTYKAKPSYPPTYKAKP 315

Qy     483 A--PTTTKEPS--PTTPKEPA-PTTTKSAPTTTKEPA-PTTTKSAPTTPKEPSPTTTKEP 536
Db     316 SYPPTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTYPSTYKAKPSYP----PTYKAKP 371

Qy     537 APTTPKEPAPTTPKKPA--PTTPKEPA--PTTPKEPAPTTTKKPAPTAPKEPAPTTPKET 592
Db     372 SYPPTYKAKPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKP-------T 424

Qy     593 APTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNT 652
Db     425 YPST-YKAKPSYP----PSYKAKPSYPPTYKAKPTYKAKPT-------YPSTYKA----K 468

Qy     653 PKEPAPTTPKEPAPTTPKE----PAPTTPKETAPTT--PKGTAPTTLKEPAPTTPKKPAP 706
Db     469 PSYPASYKAKPSYPPTYKSKSSYPSSYKPKKTYPPTYKPKLTYKPTYK-PKPSYPPSYKP 527

Qy     707 KELAPTTTKEPTSTTSDKPAPTTPKGTAPTT----PKEPAPTTPKEPAPTTPKGT---AP 759
Db     528 KTTYPPTYK-------------PKISYPPTYKAKPSYPATYKAKPSYPPTYKAKPSYPP 573

Qy     760 TTLKEPA--PTTPKKPA--PKELAPTTTKG-PTSTTSDKPAPTTPKETAPTTPKEPA--P 812
Db     574 TYKAKPSYPPTYKAKPSYKAKPTYPSTYKAKPSYPPTYKAKPSYP----PTYKAKPSYPP 629

Qy     813 TTPKKPA-PTT----PETPPPTTSEVSTPTTTKE----PTTIHKSPDESTPELSAEPTPK 863
Db     630 TYKAKPTYPSTYKAKPSYPPTYKPKISYPPTYKAKPSYPPT-YKAKPSYPPTYKAKPTYK 688

Qy     864 ALENSPK----EPGVPTT--KTPA-----ATKPEMTTTAKDKTTERDLRTTPETTTAAPK 912
Db     689 AKPTNPSTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTYPSTYKAKP- 747
```

```
RP   SEQUENCE OF 1-89 FROM N.A.
RC   TISSUE=Lung;
RX   MEDLINE=96181716; PubMed=8604237;
RA   Yu C.J., Yang P.C., Shew J.Y., Hong T.M., Yang S.C., Lee Y.C.,
RA   Lee L.N., Luh K.T., Wu C.W.;
RT   "Mucin mRNA expression in lung adenocarcinoma cell lines and
RT   tissues.";
RL   Oncology 53:118-126(1996).
RN   [12]
RP   SEQUENCE OF 1-46 FROM N.A.
RC   TISSUE=Breast carcinoma;
RA   Buluwela L., Liu Q., Luqmani Y.A., Gomm J.J., Coombes R.C.;
RL   Submitted (OCT-1992) to the EMBL/GenBank/DDBJ databases.
CC   -!- FUNCTION: DIRECT OR INDIRECT INTERACTION WITH ACTIN
CC       CYTOSKELETON.
CC   -!- SUBCELLULAR LOCATION: TYPE I MEMBRANE PROTEIN. A SECRETED FORM
CC       IS ALSO PRODUCED.
CC   -!- ALTERNATIVE PRODUCTS: VARIOUS VARIANTS ARE PRODUCED BY ALTERNATIVE
CC       SPLICING.
CC   -!- TISSUE SPECIFICITY: ABERRANTLY EXPRESSED IN HUMAN EPITHELIAL
CC       TUMORS, SUCH AS BREAST CANCER.
CC   -!- PTM: HIGHLY GLYCOSYLATED (N-AND O-LINKED CARBOHYDRATES AND SIALIC
CC       ACID).
CC   -!- POLYMORPHISM: THE NUMBER OF REPEATS IS HIGHLY POLYMORPHIC. IT
CC       VARIES FROM 21 TO 125 IN THE NORTHERN EUROPEAN POPULATION. THE
CC       MOST FREQUENT ALLELES CONTAINS 41 AND 85 REPEATS.
CC   -!- SIMILARITY: CONTAINS 1 SEA DOMAIN.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; J05582; AAA60019.1; -.
DR   EMBL; M32738; AAA35804.1; -.
DR   EMBL; M32739; AAA35806.1; -.
DR   EMBL; J05581; AAA59876.1; -.
DR   EMBL; M61170; AAB53150.1; -.
DR   EMBL; X52229; CAA36478.1; ALT_SEQ.
DR   EMBL; X52228; CAA36477.1; ALT_SEQ.
DR   EMBL; M35093; AAB59612.1; ALT_SEQ.
DR   EMBL; Z17324; CAA78972.1; -.
DR   EMBL; Z17325; CAA78973.1; -.
DR   EMBL; M31823; AAA35757.1; -.
DR   EMBL; S81781; AAD14376.1; ALT_INIT.
DR   EMBL; S81736; AAD14369.1; ALT_INIT.
DR   EMBL; M21868; AAA59874.1; ALT_SEQ.
DR   PIR; A35175; A35175.
DR   PIR; B35175; B35175.
DR   PIR; S10218; S10218.
DR   GlycoSuiteDB; P15941; -.
DR   MIM; 158340; -.
DR   MIM; 113720; -.
DR   InterPro; IPR000082; SEA.
DR   Pfam; PF01390; SEA; 1.
DR   SMART; SM00200; SEA; 1.
DR   PROSITE; PS50024; SEA; 1.
KW   Glycoprotein; Signal; Cytoskeleton; Actin-binding; Transmembrane;
KW   Repeat; Alternative splicing.
FT   SIGNAL       1     23       POTENTIAL.
FT   CHAIN       24   1255       MUCIN 1.
FT   DOMAIN      24   1162       EXTRACELLULAR (POTENTIAL).
FT   TRANSMEM  1163   1186       POTENTIAL.
FT   DOMAIN    1187   1255       CYTOPLASMIC (POTENTIAL).
FT   DOMAIN      81    960       44 X 20 AA TANDEM REPEATS.
FT   DOMAIN    1034   1151       SEA.
FT   CARBOHYD   957    957       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   975    975       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  1029   1029       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  1055   1055       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  1133   1133       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   VARSPLIC    19     19       T -> TATTAPKPAT (IN ISOFORM B).
FT   VARSPLIC    20     22       MISSING (IN ISOFORM C).
FT   VARSPLIC    20     31       MISSING (IN ISOFORM D).
FT   VARSPLIC   126    905       MISSING (IN POLYMORPHIC EPITHELIAL
FT                               ISOFORM).
FT   VARSPLIC  1077   1087       FLQIYKQGGFL -> VSIGLSFPMLP (IN SECRETED
FT                               ISOFORM).
FT   VARSPLIC  1088   1255       MISSING (IN SECRETED ISOFORM).
FT   CONFLICT     2      2       T -> A (IN REF. 11).
FT   CONFLICT   134    134       P -> Q (IN REF. 9).
FT   CONFLICT   154    154       P -> Q (IN REF. 9).
FT   CONFLICT  1021   1021       S -> T (IN REF. 3).
FT   CONFLICT  1251   1251       A -> T (IN REF. 3).
SQ   SEQUENCE  1255 AA;  122072 MW;  5E28DFC4DE7D9A82 CRC64;

  Query Match             8.8%;  Score 651;  DB 1;  Length 1255;
  Best Local Similarity   27.8%;  Pred. No. 7.1e-19;
  Matches  295;  Conservative   72;  Mismatches  460;  Indels  234;  Gaps   51;

Qy    250 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 308
Db     16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy    309 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTI---K 348
Db     66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVTSAPDNK 125

Qy    349 SAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEP------APTTPKEPAPTTTKEPAPTTTK 402
Db    126 PAPGSTAPPAHGVT-SAPDT--RPAPGSTAPPAHGVTSAPDT--RPAPGSTAPPAHGVT- 179

Qy    403 SAPTTPKEPAPTTP---------KKPAPTTPKEP------APTTPKEPTPTTP------ 440
Db    180 SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT 239

Qy    441 ----KEPAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPK----E 481
Db    240 SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT 299

Qy    482 PAPTTTKEPSPTTPKEPAPTT---TKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEP-- 536
Db    300 SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT-SAPDT--RPAPGSTAPPAH 356

Qy    537 ----APTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEP 584
Db    357 GVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAH 416

Qy    585 APTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTT 644
Db    417 GVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAP----PAHGVTSAPDTRPAPGSTA 472

Qy    645 PKA----AAPNTPKEPAPTTPK----EPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEP 696
Db    473 PPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAP----PAHGVTSAPDTRPAP 528

Qy    697 APTTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPT 744
Db    529 GSTAPPAHGVTSAPDTRPAPGSTAP-PAHGVTSAPDTRPAPGSTAPPAHGVTS----APD 583

Qy    745 TPKEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTT 790
Db    584 TRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAP-PAHGVTSAP 642

Qy    791 SDKPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVS 834
Db    643 DTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGS--TAPPAHGVTS 700

Qy    835 TPTTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEM 887
Db    701 APDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAP-----DTRPAPGSTAP-- 753
```

```
Db    612 PTPSSSTTESSSAPVT----SSTTESSSAP-VPTPSSSTTESSSAPVPTPSSSTTESSSA 666

Qy    640 PAP-----TTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPT-TPKGTAPTTL 693

Db    667 PVPTPSSSTTESSSAPVT----SSTTESSSAPVT----SSTTESSSAPVPTPSSSTTESS 718

Qy    694 KEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAP---TTPKEPA 750

Db    719 SAPVPTPSSSTTESSSAPVPTPS-SSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSSA 777

Qy    751 PT-TPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKE 809

Db    778 PVPTPSSSTTESSSAPVPTPSSSTTESSVAPVPTPSSSSNITSSAPSSTPFSSSTESSSV 837

Qy    810 PAP-----TTPKKPAPTTPETPPPTTSEVSTPT-----TTKEPTTIHKSPDESTPELSAE 859

Db    838 PVPTPSSSTTESSSAPVSSSTTESSVAPVPTPSSSSNITSSAPSSI---PFSSTTESFST 894

Qy    860 PTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETAT 919

Db    895 GT-TVTPSSSKYPGSQTETSVSSTTETTIVPTKTTTSVTTPSTTTITTTVCSTGTNSAGE 953

Qy    920 TTEKTTESKITAT--TTQVTSTTTQDTTPFKITTLKTTTLAPKVTT---TKKTITTT-EI 973

Db    954 TTSGCSPKTVTTTVPTTTTTSVTTSSTTTITTTVCSTGTNSAGETTSGCSPKTITTTVPC 1013

Qy    974 MNKPEETAKPKDRATNSKATTPKPQKP----------------TKAPKKPTSTKKPKTM 1016

Db   1014 STSPSETA--------SESTTTSPTTPVTTVVSTTVVTTEYSTSTKPGGEITTTFVTKNI 1065

Qy   1017 PRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGA 1076

Db   1066 PTTYLTTIAPTP--SVTTVTNFTPTT-ITTTVCSTGT----------NSAGETTSGC 1109

Qy   1077 EGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSDETNICNGKPVDGLTTLRNG 1136

Db   1110 SPKTVTTTVPCSTGTGEYTTEATTLVTTAVTTTVVTTESSTGTNSA-GKTTTGYTTKSVP 1168

Qy   1137 TLVAFRGHYFWMLSPFSPPSPARRITEVWGIPSPIDTVFTRC----NCEGKT 1184

Db   1169 TT------YVTTLAPSAPVTPATN-----AVPTTITT--TECSAATNAAGET 1207
```

```
RESULT   4
MUC1_HUMAN
ID   MUC1_HUMAN     STANDARD;      PRT;  1255 AA.
AC   P15941; P15942; P13931; P17626; Q14128; Q16442; Q16437; Q9Y4J2;
DT   01-JAN-1990 (Rel. 13, Created)
DT   01-APR-1990 (Rel. 14, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   MUCIN 1 PRECURSOR (POLYMORPHIC EPITHELIAL MUCIN) (PEM) (PEMT)
DE   (EPISIALIN) (TUMOR-ASSOCIATED MUCIN) (CARCINOMA-ASSOCIATED MUCIN)
DE   (TUMOR-ASSOCIATED EPITHELIAL MEMBRANE ANTIGEN) (EMA) (H23AG) (PEANUT-
DE   REACTIVE URINARY MUCIN) (PUM) (BREAST CARCINOMA-ASSOCIATED ANTIGEN
DE   DF3).
GN   MUC1.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Pancreas;
RX   MEDLINE=90368716; PubMed=2394722;
RA   Lan M.S., Batra S.K., Qi W.-N., Metzgar R.S., Hollingsworth M.A.;
RT   "Cloning and sequencing of a human pancreatic tumor mucin cDNA.";
RL   J. Biol. Chem. 265:15294-15299(1990).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=90202794; PubMed=2318825;
RA   Ligtenberg M.J.L., Vos H.L., Gennissen A.M.C., Hilkens J.;
RT   "Episialin, a carcinoma-associated mucin, is generated by a
RT   polymorphic gene encoding splice variants with alternative amino
RT   termini.";
RL   J. Biol. Chem. 265:5573-5578(1990).
RN   [3]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Breast carcinoma;
RX   MEDLINE=90368715; PubMed=1697589;
RA   Gendler S.J., Lancaster C.A., Taylor-Papadimitriou J., Duhig T.,
RA   Peat N., Burchell J., Pemberton L., Lalani E.-N., Wilson D.;
RT   "Molecular cloning and expression of human tumor-associated
RT   polymorphic epithelial mucin.";
RL   J. Biol. Chem. 265:15286-15293(1990).
RN   [4]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=91097524; PubMed=2268309;
RA   Lancaster C.A., Peat N., Duhig T., Wilson D.,
RA   Taylor-Papadimitriou J., Gendler S.J.;
RT   "Structure and expression of the human polymorphic epithelial mucin
RT   gene: an expressed VNTR unit.";
RL   Biochem. Biophys. Res. Commun. 173:1019-1029(1990).
RN   [5]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Breast carcinoma;
RX   MEDLINE=90276413; PubMed=2351132;
RA   Wreschener D.H., Hareuveni M., Tsarfaty I., Smorodinsky N.,
RA   Horov J., Zaretsky J., Kotkes P., Weiss M., Lathe R., Dion A.,
RA   Keydar I.;
RT   "Human epithelial tumor antigen cDNA sequences. Differential splicing
RT   may generate multiple protein forms.";
RL   Eur. J. Biochem. 189:463-473(1990).
RN   [6]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Breast carcinoma;
RX   MEDLINE=90276414; PubMed=2112460;
RA   Hareuveni M., Tsarfaty I., Zaretsky J., Kotkes P., Horev J.,
RA   Zrihan S., Weiss M., Green S., Lathe R., Keydar I., Wreschner D.H.;
RT   "A transcribed gene, containing a variable number of tandem repeats,
RT   codes for a human epithelial tumor antigen. cDNA cloning, expression
RT   of the transfected gene and over-expression in breast cancer
RT   tissue.";
RL   Eur. J. Biochem. 189:475-486(1990).
RN   [7]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=91033045; PubMed=1688329;
RA   Tsarfaty I., Hareuveni M., Horev J., Zaretsky J., Weiss M.,
RA   Jeltsch J.M., Garnier J.M., Lathe R., Keydar I., Wreschner D.H.;
RT   "Isolation and characterization of an expressed hypervariable gene
RT   coding for a breast-cancer-associated antigen.";
RL   Gene 93:313-318(1990).
RN   [8]
RP   PARTIAL SEQUENCE FROM N.A.
RX   MEDLINE=88330762; PubMed=3417635;
RA   Gendler S.J., Taylor-Papadimitriou J., Duhig T., Rothbard J.,
RA   Burchell J.;
RT   "A highly immunogenic region of a human polymorphic epithelial mucin
RT   expressed by carcinomas is made up of tandem repeats.";
RL   J. Biol. Chem. 263:12820-12823(1988).
RN   [9]
RP   SEQUENCE OF 1-169 FROM N.A.
RX   MEDLINE=90088473; PubMed=2597151;
RA   Abe M., Siddiqui J., Kufe D.;
RT   "Sequence analysis of the 5' region of the human DF3 breast
RT   carcinoma-associated antigen gene.";
RL   Biochem. Biophys. Res. Commun. 165:644-649(1989).
RN   [10]
RP   SEQUENCE OF 1-109 FROM N.A.
RC   TISSUE=Thyroid;
RX   MEDLINE=96183746; PubMed=8608966;
RA   Weiss M., Baruch A., Keydar I., Wreschner D.H.;
RT   "Preoperative diagnosis of thyroid papillary carcinoma by reverse
RT   transcriptase polymerase chain reaction of the MUC1 gene.";
RL   Int. J. Cancer 66:55-59(1996).
RN   [11]
```

```
Qy     810 PAPTTPKKPAPT-TPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENS 868
           | |:   :| |: ||| | || :    || : |||      || ||  | ||||
Db    1254 PTPS--DEPTPSETPEEPIPTDTPSDEPTPSDEPT----PSDEPTP--SDEPTPSDEPTP 1305

Qy     869 PKEPGVPTTKTPAAT--KPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTE 926
             ||      ||: |  :|  | |  |: |  |      | :  |  : |   : | |
Db    1306 SDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSD----EPTPSDEPTPSDEPTPSDEPTPS 1361

Qy     927 SKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDR 986
           : | :     | | :: ||       |||  |  :||                       |
Db    1362 DEPTPSDEPTPSETPEEPTP-------TTTPTPTPSTT-----------------PTSG 1396

Qy     987 ATNSKAT--------TPKPQKPTKAP-KKPTSTKKPKTMPRVRKPKTTPTPRKM------ 1031
           : |  :           |  ||  |  |||||  |     : :|  :  |  :
Db    1397 SGGSGGSGGGGGGGGGTVPTSPTPTPTSKPTSTPAP---TEIEEPTPSDVPGAIGGEHRA 1453

Qy    1032 -------TSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHM- 1083
                |  || | |   |  :            :||:     :::: ||:  :|:
Db    1454 YLRGYPDGSFRPERNITRAEAAVIF----------AKLL----GADESYGAQSASPYSD 1498

Qy    1084 LLRPH--------------------VFMPE--------VTPDMDYLPRVPNQGI----- 1109
           |   |                    | |:         |  : :| :|  | |
Db    1499 LADTHWAAWAIKFATSQGLFKGYPDGTFKPDQNITRAEFATVVLHFLTKVKGQEIMSKLA 1558

Qy    1110 ---IINPMLSDETNICNG----KPVDGLTTL------RNGTLVAFRGHYFWMLSPFSPPS 1156
             | ||   |    | |    : :: ||:|       :||             | | :
Db    1559 TIDISNPKFDD----CVGHWAQEFIEKLTSLGYISGYPDGT-------------FKPQN 1600

Qy    1157 PARRITEVWGIPSPIDTVFTRCNCEGKTFFFKD--SQYWRFTNDIKD 1201
           :|  |   : :|:    |    |    | |    || |  || |
Db    1601 YIKRSESV----ALINRALERGPLNGAPKLFPDVNESYWAF-GDIMD 1642
```

```
RESULT   3
AMYH_YEAST
ID   AMYH_YEAST     STANDARD;      PRT;  1367 AA.
AC   P08640; P08068;
DT   01-AUG-1988 (Rel. 08, Created)
DT   01-FEB-1995 (Rel. 31, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   GLUCOAMYLASE S1/S2 PRECURSOR (EC 3.2.1.3) (GLUCAN 1,4-ALPHA-
DE   GLUCOSIDASE) (1,4-ALPHA-D-GLUCAN GLUCOHYDROLASE).
GN   STA1 OR STA2 OR MAL5 OR YIR019C.
OS   Saccharomyces cerevisiae (Baker's yeast).
OC   Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
OC   Saccharomycetales; Saccharomycetaceae; Saccharomyces.
OX   NCBI_TaxID=4932;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=S288C / AB972;
RA   Barrell B.G., Badcock K., Bankier A.T., Bowman S., Brown D.,
RA   Churcher C.M., Connor R., Copsey T., Dear S., Devlin K., Fraser A.,
RA   Gentles S., Hamlyn N., Horsnell T.S., Hunt S., Jagels K., Jones M.,
RA   Louis E., Lye G., Moule S., Moule T., Odell C., Pearson D.,
RA   Rajandream M.A., Riles L., Rowley N., Skelton J., Smith V.,
RA   Walsh S.V., Whitehead S.;
RL   Submitted (DEC-1994) to the EMBL/GenBank/DDBJ databases.
RN   [2]
RP   SEQUENCE OF 1-242 AND 762-1331 FROM N.A.
RX   MEDLINE=87194600; PubMed=3106330;
RA   Yamashita I., Nakamura M., Fukui S.;
RT   "Gene fusion is a possible mechanism underlying the evolution of
RT   STA1.";
RL   J. Bacteriol. 169:2142-2149(1987).
RN   [3]
RP   SEQUENCE OF 1-31 FROM N.A.
RC   STRAIN=SPX101-1C;
RX   MEDLINE=89031230; PubMed=3141213;
RA   Pardo J.M., Ianez E., Zalacain M., Claros M.G., Jimenez A.;
RT   "Similar short elements in the 5' regions of the STA2 and SGA genes
RT   from Saccharomyces cerevisiae.";
RL   FEBS Lett. 239:179-184(1988).
```

```
CC   -!- CATALYTIC ACTIVITY: HYDROLYSIS OF TERMINAL 1,4-LINKED ALPHA-D-
CC       GLUCOSE RESIDUES SUCCESSIVELY FROM NON-REDUCING ENDS OF THE CHAINS
CC       WITH RELEASE OF BETA-D-GLUCOSE.
CC   -!- SIMILARITY: TO S.POMBE SPBC215.13.
CC   -!- SIMILARITY: SOME, TO S.POMBE SPCC285.13C.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; Z38061; CAA86176.1; -.
DR   EMBL; M16164; AAA35014.1; -.
DR   EMBL; M16165; AAA35015.1; -.
DR   EMBL; X13857; CAA32069.1; -.
DR   PIR; B26877; B26877.
DR   PIR; A26877; A26877.
DR   PIR; S48478; S48478.
DR   SGD; S0001458; MUC1.
KW   Hydrolase; Glycosidase; Polysaccharide degradation; Glycoprotein;
KW   Signal; Multigene family.
FT   SIGNAL       1     21       POTENTIAL.
FT   CHAIN       22   1367       GLUCOAMYLASE S1/S2.
FT   DOMAIN     210   1367       SER/THR-RICH.
FT   CARBOHYD   817    817       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   874    874       N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE   1367 AA;  136110 MW;  91C00E2DBD61AA9D CRC64;
```

```
  Query Match             10.9%;  Score 809;  DB 1;  Length 1367;
  Best Local Similarity   27.5%;  Pred. No. 6.1e-25;
  Matches  328;  Conservative  112;  Mismatches  526;  Indels  226;  Gaps   53;

Qy      61 DCDAQCKKYDK---CCPDYESFCAEVHNPTSPPSSKKAPPPSGASQTIKSTTKRSPKPPN 117
           |    |  ||       |: |   :    :  :| :   |  |::  |||
Db     174 DLSTGCNNYDNQGHSQTDFPGFYWNIDCDNNCGGTKSSTTTSSTSES--STT-------- 223

Qy     118 KKKTKKVIESEEITEEHSVSENQESSSSSSSSSSSSTIWKIKSSKNSAANRELQKKLKVK 177
              |    ||   |   | |    ||:| ||:|||:|      :  :             |:
Db     224 ---TSSTSESSTTTSSTSESSTTTSSTSESSTSSSTT------APATPTTTSCTKEKPTP 274

Qy     178 DNKKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPR 237
           : ||:|||  ||  |                 |||  |      | :|: |  | |   |
Db     275 PTTTSCTKEKPT--PPHHD------------TTPCT-----KKKTTTSKTCTKKTTTPV 314

Qy     238 PSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDL 297
           |:  |:| |:: :|        | |  ::||          :::||  | |  ::|:
Db     315 PT--PSSSTTESSS------APVPTPSSSTT----------ESSSAPVTSSTTESSS--- 353

Qy     298 APTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEP 357
           ||       ||| : ||     |  |       ::    :|||      |    |: ||
Db     354 APV------PTPSSSTTESSSAPVTSSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSS 407

Qy     358 APTT-----TKSAP----TTPKEPAPTTTKEPAPTTPKEPAP----TTTKEPAPTTTKSA 404
           || |     : |||     ||     || |:     ||    ||    ||     ||  | |:
Db     408 APVTSSTTESSSAPVTSSTTESSSAPVTSS----TTESSSAPVTSSTTESSSAPVPTPSS 463

Qy     405 PTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTP---TTPKEPAPTTKEPAPTTPKEPAPT 461
           ||    || |    : ||     ||      |||   ||    || |   : ||    ||
Db     464 STTESSSAPVT----SSTTESSSAPV----PTPSSSTTESSSAPVT---SSTTESSSAPV 512

Qy     462 APKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSA 521
                 ||      :  |||| :   | ||    || |:    ||    ||  | |:
Db     513 -------PTPSSSTTESSSAPAPTPS---SSTTESSSAPVTSS---TTESSSAPVPTPSS 559

Qy     522 PTTPKEPSPTTTKEPAPTTPKEPAPT-TPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTA 580
           ||    :| |:     ||    ||  ||       :   | ||    |: :||:  :  |
Db     560 STTESSSTPVTSS----TTESSSAPVPTPSSSTTESSSAPVPT----PSSSTTESSSAPA 611

Qy     581 PKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPT-TPEEPTPTTPEE 639
```

```
Qy     523 --TTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTA 580
           |||  |  |||  |: ||   | |||   |:|||   |  ||     ||||  |  |
Db    1701 PTTTPSSPI-TTTTTPSSTTTPSPPPTTMTTPSPTTTPSPPTTTMTTLPPTTTSSPLTTT 1759

Qy     581 PKEPAPTTP-----KETAPTTP------------------------------------ 597
           |  |: | |       | ||||
Db    1760 PLPPSITPPTFSPFSTTTPTTPCVPLCNWTGWLDSGKPNFHKPGGDTELIGDVCGPGWAA 1819

Qy     598 -----KKLTP---------------------------------------------- 602
                : |
Db    1820 NISCRATMYPDVPIGQLGQTVVCDVSVGLICKNEDQKPGGVIPMAFCLNYEINVQCCECV 1879

Qy     603 TTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPT-TPEEPAPTTPKAAAPNTPKEPAPTTP 661
           | |  :  || | | | |   :  ||   |||| |  :   |||          | ||
Db    1880 TQPTTMTTTTTENPTPPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGT 1939

Qy     662 KEP-----APTTPKEPAPT-TPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTK 715
           : |       ||   | || |  :|  |||  |  |    | ||  : |    :  |||
Db    1940 QTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTV 1999

Qy     716 EPTSTTSDKPAP-TTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPA 774
           || | :    | |||  |  |    | ||  : |  |||  |  |    | ||  : |
Db    2000 TPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPT-TTPITTTTTVTPTPTPTGTQTPT 2058

Qy     775 PKELAPTTTKGPTSTTSDKPAP-TTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEV 833
           :  |||  || | :    | |||  |  |    | ||  : |  ||| |   | :
Db    2059 TTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPT-TTPITTTTTVTPT 2117

Qy     834 STPTTTKEPTTIHKSPDESTPELSAEPTPKALENSPKEP----------GVPT-TKTPAA 882
           |||  |: |||   :|  :|  ::  |||   :    |            || |:||
Db    2118 PTPTGTQTPTT---TPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTT 2174

Qy     883 TKPEMTTTAKDKTTERDLR---TTP--ETTTAAPKMT---KETATTTEKTTESKITATTT 934
           |    |||     |    :   |||   |||  |  |     :| |||  || : :| | |
Db    2175 TPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPT 2234

Qy     935 QVTSTTTQDTTPFKITTLKTTTLAPKVTTT-KKTITTTEIMNKPEETAKPKDRATNSKAT 993
             | |  |||  |||  |||: |  | |  :| ||| |     |  |    | :  |
Db    2235 P-TGTQTPTTTP--ITT--TTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTT 2289

Qy     994 TPKPQKPTKAP-KKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNP---TSRIAEAML 1049
           ||     |  |   || |: | | |       ||||    :  |  |   |: :
Db    2290 TPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPT 2349

Qy    1050 QTTTR-PNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTP 1096
           | |: |  || :    | |     |    :||     |      |||
Db    2350 PTGTQTPTTTPITTTTTVTPTPTPTG---TQTP--TTTPITTTTTVTP 2392


RESULT   2
SLP1_CLOTM
ID   SLP1_CLOTM     STANDARD;      PRT;  1664 AA.
AC   Q06852;
DT   01-JUN-1994 (Rel. 29, Created)
DT   01-JUN-1994 (Rel. 29, Last sequence update)
DT   01-FEB-1996 (Rel. 33, Last annotation update)
DE   CELL SURFACE GLYCOPROTEIN 1 PRECURSOR (OUTER LAYER PROTEIN B) (S-LAYER
DE   PROTEIN 1).
GN   OLPB.
OS   Clostridium thermocellum.
OC   Bacteria; Firmicutes; Bacillus/Clostridium group; Clostridiaceae;
OC   Clostridium.
OX   NCBI_TaxID=1515;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=NCIB 10682;
RX   MEDLINE=93209931; PubMed=8458832;
RA   Fujino T., Beguin P., Aubert J.-P.;
RT   "Organization of a Clostridium thermocellum gene cluster encoding the
RT   cellulosomal scaffolding protein CipA and a protein possibly involved
RT   in attachment of the cellulosome to the cell surface.";
RL   J. Bacteriol. 175:1891-1899(1993).
CC   -!- SUBUNIT: ASSEMBLED INTO MONO-LAYERED CRYSTALLINE ARRAYS.
CC   -!- SUBCELLULAR LOCATION: CELL WALL.
CC   -!- SIMILARITY: CONTAINS 4 S-LAYER HOMOLOGY (SLH) DOMAINS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X67506; CAA47841.1; -.
DR   InterPro; IPR001119; SLH.
DR   Pfam; PF00395; SLH; 3.
DR   PROSITE; PS01072; SLH_DOMAIN; 2.
KW   Cell wall; S-layer; Signal; Repeat.
FT   SIGNAL       1     28       POTENTIAL.
FT   CHAIN       29   1664       CELL SURFACE GLYCOPROTEIN 1.
FT   DOMAIN      36    763       4 X 156 AA APPROXIMATE REPEATS.
FT   REPEAT      36    191       1.
FT   REPEAT     207    363       2.
FT   REPEAT     409    565       3.
FT   REPEAT     607    763       4.
FT   DOMAIN     771   1377       APPROXIMATE TANDEM REPEATS OF
FT                               T-P-S-D-E-P.
FT   DOMAIN    1378   1449       GLY/PRO/SER/THR-RICH.
FT   DOMAIN    1453   1494       SLH 1 (INCOMPLETE).
FT   DOMAIN    1495   1565       SLH 2.
FT   DOMAIN    1566   1625       SLH 3.
FT   DOMAIN    1626   1646       SLH 4 (INCOMPLETE).
SQ   SEQUENCE   1664 AA;  178194 MW;  5F396695BA9FE74B CRC64;


  Query Match             12.8%;  Score 950;  DB 1;  Length 1664;
  Best Local Similarity   31.0%;  Pred. No. 2.6e-30;
  Matches  312;  Conservative  104;  Mismatches  361;  Indels  230;  Gaps   55;

Qy     303 VLAKPTP-KAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTT 361
           |: :| | || :    |  |   ||||:          ||||:    ||   || |:
Db     758 VVIQPAPIKAASDEPIPTDTPSDEPTPS----------DEPTPS---DEPTPSDEPTPSD 804

Qy     362 TKSAPTTPKEPAPTTTKEPAPTTPKEPA----PTTTKEPAPTTTKSAPTTPKEPAPTTPK 417
             :  ||:|| || |    ||   ||     || : || |:   :   ||:|| ||
Db     805 EPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTP 864

Qy     418 KPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPT-TPKEPAPTAPKKPAPTTPKEPAP 476
           ||   || |:   ||||:   || | : || |: ||:|| ||      ||   || |
Db     865 SDEPTPSDEPTPS--DEPTPS--DEPTP-SDEPTPSETPEEPIPTDTPSDEPTPSDEPTP 919

Qy     477 TTPKEPA--PTTTKEPSPT-TPKEPAPTTTKS-APTTTKEPAPT---TTKSAPTTPKEPS 529
           :   |:  || : ||:|: ||:|| || | |  || : || |:   |    ||   ||:
Db     920 SDEPTPSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPT 979

Qy     530 ----PTTTKEPAPT-TPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEP 584
               || : || |: ||:|| ||      ||   || |:    |:   |   ||   ||
Db     980 PSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSDEP 1039

Qy     585 APT---TPKETAPTTPKKLTPTTPEKLAPTTPEKPAPT---TPEELAPTTPEEPTPT-TP 637
           |:   || |    ||::  ||      ||  ::| |:   || :  ||  :|||| ||
Db    1040 TPSDEPTPSE----TPEEPIPTDTPSDEPTPSDEPTPSD-EPTPSDEPTPSETP 1094

Qy     638 EEPAPTTPKAAAPNTPKEPAPTTPKEPA--PTTPKEPAPT-TPKETAPTTPKGTAPTTLK 694
           ||| ||   :  |    || |:    |:  ||   || |: ||:|  ||        ||
Db    1095 EEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSD 1154

Qy     695 EPAPTTPKKPAPKELAPTTTKEPT-STTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTT 753
           || |:   :| | :  || : ||| | | ::| ||      ||   || |:   || |:
Db    1155 EPTPS--DEPTPSD-EPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPS--DEPTPS- 1208

Qy     754 PKGTAPTTLKEPAPT-TPKKPAPKELAPTTTKGPTSTTSDKPAPT---TPKETAPTTPKE 809
             ||   || |: ||::|           || | ||:| |:   || :  ||   |
Db    1209 ---DEPTPSDEPTPSETPEEPI----------PTDTPSDEPTPSDEPTPSD-EPTPSDE 1253
```

```
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------------
DR   EMBL; L21998; AAB95295.1; -.
DR   EMBL; M74027; AAA59875.1; -.
DR   EMBL; M94131; AAA59163.1; -.
DR   EMBL; M94132; AAA59164.1; -.
DR   MIM; 158370; -.
DR   InterPro; IPR000359; Cys_knot.
DR   InterPro; IPR000561; EGF-like.
DR   InterPro; IPR002400; GF_cysknot.
DR   InterPro; IPR001007; VWFC.
DR   InterPro; IPR001846; Vwd.
DR   Pfam; PF00007; Cys_knot; 1.
DR   Pfam; PF00094; vwd; 4.
DR   PRINTS; PR00438; GFCYSKNOT.
DR   SMART; SM00214; VWC; 2.
DR   SMART; SM00011; VWC_def; 2.
DR   SMART; SM00216; VWD; 4.
DR   PROSITE; PS00022; EGF_1; UNKNOWN_1.
DR   PROSITE; PS01185; CTCK_1; 1.
DR   PROSITE; PS01225; CTCK_2; 1.
DR   PROSITE; PS01208; VWFC; 2.
KW   Glycoprotein; Repeat; Signal.
FT   SIGNAL       1     20       POTENTIAL.
FT   CHAIN       21   5179       MUCIN 2.
FT   DOMAIN    1401   1747       APPROXIMATE REPEATS.
FT   REPEAT    1401   1416       1.
FT   REPEAT    1417   1432       2.
FT   REPEAT    1433   1448       3.
FT   REPEAT    1449   1464       4.
FT   REPEAT    1465   1471       5.
FT   REPEAT    1472   1478       6.
FT   REPEAT    1479   1494       7A.
FT   REPEAT    1495   1517       7B.
FT   REPEAT    1518   1533       8A.
FT   REPEAT    1534   1556       8B.
FT   REPEAT    1557   1572       9A.
FT   REPEAT    1573   1596       9B.
FT   REPEAT    1597   1612       10A.
FT   REPEAT    1613   1635       10B.
FT   REPEAT    1636   1651       11A.
FT   REPEAT    1652   1675       11B.
FT   REPEAT    1676   1683       12.
FT   REPEAT    1684   1699       13.
FT   REPEAT    1700   1715       14.
FT   REPEAT    1716   1731       15.
FT   REPEAT    1732   1747       16.
FT   DOMAIN    4815   4886       VWFC 1.
FT   DOMAIN    4924   4991       VWFC 2.
FT   DOMAIN    5075   5160       CTCK.
FT   DISULFID  5075   5122       BY SIMILARITY.
FT   DISULFID  5089   5136       BY SIMILARITY.
FT   DISULFID  5098   5152       BY SIMILARITY.
FT   DISULFID  5102   5154       BY SIMILARITY.
FT   DISULFID     ?   5159       BY SIMILARITY.
FT   CARBOHYD   163    163       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   423    423       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   670    670       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   770    770       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   894    894       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  1139   1139       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  1154   1154       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  1215   1215       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  1230   1230       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  1246   1246       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  1787   1787       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  1820   1820       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  4339   4339       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  4351   4351       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  4362   4362       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  4373   4373       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  4422   4422       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  4438   4438       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  4502   4502       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  4616   4616       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  4627   4627       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  4752   4752       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  4787   4787       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  4881   4881       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  4888   4888       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  4955   4955       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  4970   4970       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  5019   5019       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  5038   5038       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  5069   5069       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CONFLICT  1351   1351       H -> L (IN REF. 3).
FT   CONFLICT  1412   1412       T -> S (IN REF. 3).
FT   CONFLICT  1449   1449       L -> P (IN REF. 3).
FT   CONFLICT  1504   1504       M -> T (IN REF. 3).
FT   CONFLICT  4192   4192       G -> S (IN REF. 2).
SQ   SEQUENCE  5179 AA;  540295 MW;  85CD7571FB9A5663 CRC64;
```

```
  Query Match           16.0%;  Score 1188;  DB 1;  Length 5179;
  Best Local Similarity 29.5%;  Pred. No. 4.2e-39;
  Matches  403;  Conservative   85;  Mismatches  536;  Indels  344;  Gaps   53;

Qy      19 TCNCDY--NCQHYMECCPDFKRVCTAELSCKGRCFESFERGRECDCDAQCKKYDKCCPDY 76
           :|:||   :|: :      : : || | :|       | |  :  |  |  |:   | :
Db    1079 SCSCDTGGDCECFCSAVASYAQECTKEGAC------VFWRTPDL-CPIFCDYYN---PPH 1128

Qy      77 ESFCAEVHNPTSPPSSKKAPPPSGASQTI-----KSTTKRSPK--PPNKKKTKKVIESEE 129
           |  |   : |    | :     :|    |     :    | ||  |  ::  || : :::
Db    1129 E--CEWHYEPCGNRSFETCRTINGIHSNISVSYLEGCYPRCPKDRPIYEEDLKKCVTADK 1186

Qy     130 ----ITEEH----SVSENQESSSSSSSSSSSTI------------------WKIKSSK 162
               : |   :    :|:  |   ::||  :                    |:|
Db    1187 CGCYVEDTHYPPGASVPTEETCKSCVCTNSSQVVCRPEEGKILNQTQDGAFCYWEI---- 1242

Qy     163 NSAANRELQK--------------------KLKVKDNKKNRTKKKPTPKPPVV------- 195
           |  ::|                        |     |    ||    |
Db    1243 -CGPNGTVEKHFNICSITTRPSTLTTFTTITLPTTPTSFTTTTTTTTPTSSTVLSTTPKL 1301

Qy     196 ----------DEAGSGLDNGDFK----------------VTTPDTSTTQH-NKVSTSPK- 227
                    |   || |:|| :                  |  |  |  ||  ||
Db    1302 CCLWSDWINEDHPSSGSDDGDREPFDGVCGAPEDIECRSVKDPHLSLEQHGQKVQCDVSV 1361

Qy     228 ---------------------------------ITTAKPINPRPSLPPNSDTSKETSLT 253
                                               |||  |    || || : |:   : |
Db    1362 GFICKNEDQFGNGPFGLCYDYKIRVNCCWPMDKCITTPSPPTTTPSPPPTTTTTLPPTTT 1421

Qy     254 VNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAET 313
             :  ||      ||||  |:|    ||:                  ||:     |:|   |
Db    1422 PSPPTT-----TTTTPPPTTTPSPPITTTTTP-----------LPTT----TPSPPIST 1460

Qy     314 TTKGPALTTPKEPT-----PTTPKEPASTTPKEPTPTTIKSAP-TTPKEP-------APT 360
           ||  |  |||  ||     |||   | :||   | |||  | | |||  |         ||
Db    1461 TTTPPPTTTPSPPTTTPSPPTTTPSPPTTTTTTPPPTTPSPPMTTPITPPASTTTLPPT 1520

Qy     361 TTKSAP-TTPKEPAPTTTKEPAPTTPKEP-------APTTTKEPAPTTTKSAP--TTPKE 410
           || | | ||   | ||||  |  |||  |         ||||  | |||| : |  |||
Db    1521 TTPSPPTTTTTTPPPTTTPSPPTTTPITPPTSTTTLPPTTTPSPPPTTTTTPPPTTTPSP 1580

Qy     411 PAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEP-------APTAP 463
           |  |||  |  ||   |  |||  || ||   |  ||  |  |||  |         ||
Db    1581 PTTTTPSPPTITTTPPPTTTPSPPTTTTTTPPPTTTPSPPTTTPITPPTSTTTLPPTTT 1640

Qy     464 KKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAP- 522
           | |||   | |||   |  |||  |  ||   | |||| |:| ||    |||| : |
Db    1641 PSPPPTTTTTPPPTTTPSPPTTTTPSPPITTTTTPPPTTTPSSPITTTPSPPTTTMTTPS 1700
```

GenCore version 4.5
Copyright (c) 1993 - 2000 Compugen Ltd.

OM protein - protein search, using sw model

Run on: April 26, 2002, 16:19:03 ; Search time 47.4 Seconds
(without alignments)
1067.457 Million cell updates/sec

Title: US-09-556-246-1_COPY_25_1404
Perfect score: 7410
Sequence: 1 QDLSSCAGRCGEGYSRDATC.....ARAITTRSGQTLSKVWYNCP 1380

Scoring table: BLOSUM62
Gapop 10.0 , Gapext 0.5

Searched: 100059 seqs, 36664827 residues

Total number of hits satisfying chosen parameters: 100059

Minimum DB seq length: 0
Maximum DB seq length: 2000000000

Post-processing: Minimum Match 0%
Maximum Match 100%
Listing first 45 summaries

Database : SwissProt_39:*

Pred. No. is the number of results predicted by chance to have a score greater than or equal to the score of the result being printed, and is derived by analysis of the total score distribution.

SUMMARIES

Result No.	Score	% Query Match	Length	DB	ID	Description
1	1188	16.0	5179	1	MUC2_HUMAN	Q02817 homo sapien
2	950	12.8	1664	1	SLP1_CLOTM	Q06852 clostridium
3	809	10.9	1367	1	AMYH_YEAST	P08640 saccharomyc
4	651	8.8	1255	1	MUC1_HUMAN	P15941 h mucin 1 p
5	559	7.5	875	1	FP1_MYTED	Q25460 mytilus edu
6	555.5	7.5	2700	1	ZAN_HUMAN	Q9y493 homo sapien
7	551	7.4	620	1	EXTN_TOBAC	P13983 nicotiana t
8	551	7.4	1087	1	NFH_MOUSE	P19246 mus musculu
9	530.5	7.2	1162	1	TCNA_TRYCR	P23253 trypanosoma
10	530	7.2	865	1	CPN_DROME	Q02910 drosophila
11	518.5	7.0	872	1	FP1_MYTCO	Q25434 mytilus cor
12	517.5	7.0	831	1	NFH_RAT	P16884 rattus norv
13	509.5	6.9	662	1	MUC1_XENLA	Q05049 xenopus lae
14	503.5	6.8	1970	1	RPB1_HUMAN	P24928 homo sapien
15	502	6.8	1020	1	NFH_HUMAN	P12036 homo sapien
16	498.5	6.7	1970	1	RPB1_MOUSE	P08775 mus musculu
17	493.5	6.7	467	1	RPB1_CRIGR	P11414 cricetulus
18	490	6.6	2142	1	BAT2_HUMAN	P48634 homo sapien
19	488.5	6.6	826	1	SSP2_PLAYO	Q01443 plasmodium
20	475.5	6.4	267	1	EXTN_MAIZE	P14918 zea mays (m
21	471.5	6.4	5376	1	ZAN_MOUSE	O88799 mus musculu
22	468.5	6.3	634	1	HWP1_CANAL	P46593 candida alb
23	467	6.3	817	1	VRP1_YEAST	P37370 saccharomyc
24	454	6.1	797	1	VGLX_HSVEB	P28968 equine herp
25	454	6.1	1161	1	YJ9P_YEAST	P47179 saccharomyc
26	453.5	6.1	1794	1	YAV1_SCHPO	Q10172 schizosacch
27	448.5	6.1	670	1	VG50_HSVI1	Q00130 ictalurid h
28	446	6.0	1229	1	N121_HUMAN	Q9y2n3 homo sapien
29	444.5	6.0	2774	1	MAPA_RAT	P34926 rattus norv
30	442.5	6.0	751	1	FP1_MYTGA	Q27409 mytilus gal
31	439.5	5.9	1083	1	T2D3_HUMAN	O00268 homo sapien
32	436.5	5.9	1411	1	TCOF_HUMAN	Q13428 homo sapien
33	434.5	5.9	3256	1	KI67_HUMAN	P46013 homo sapien
34	433.5	5.9	3164	1	TEGU_HSV11	P10220 herpes simp
35	432.5	5.8	439	1	XP2_XENLA	P17437 xenopus lae
36	428	5.8	2517	1	NCR2_HUMAN	Q9y618 h nuclear r
37	427.5	5.8	1251	1	YQU3_CAEEL	Q09550 caenorhabdi
38	426	5.7	2715	1	TRX2_HUMAN	Q9umn6 homo sapien
39	425.5	5.7	1185	1	DRPL_HUMAN	P54259 homo sapien
40	424.5	5.7	1125	1	MAP4_MOUSE	P27546 mus musculu
41	424	5.7	307	1	SGS3_DROME	P02840 drosophila
42	423.5	5.7	3421	1	TEGU_HSVEB	P28955 equine herp
43	422.5	5.7	1183	1	DRPL_RAT	P54258 rattus norv
44	419.5	5.7	2476	1	ZAN_PIG	Q28983 sus scrofa
45	411.5	5.6	907	1	VGP3_EBV	P03200 epstein-bar

ALIGNMENTS

RESULT 1
MUC2_HUMAN
ID MUC2_HUMAN STANDARD; PRT; 5179 AA.
AC Q02817; Q14878;
DT 01-JUN-1994 (Rel. 29, Created)
DT 01-NOV-1997 (Rel. 35, Last sequence update)
DT 20-AUG-2001 (Rel. 40, Last annotation update)
DE MUCIN 2 PRECURSOR (INTESTINAL MUCIN 2).
GN MUC2 OR SMUC.
OS Homo sapiens (Human).
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX NCBI_TaxID=9606;
RN [1]
RP SEQUENCE FROM N.A.
RC TISSUE=Intestine;
RX MEDLINE=94132002; PubMed=8300571;
RA Gum J.R. Jr., Hicks J.W., Toribara N.W., Siddiki B., Kim Y.S.;
RT "Molecular cloning of human intestinal mucin (MUC2) cDNA.
RT Identification of the amino terminus and overall sequence similarity
RT to prepro-von Willebrand factor.";
RL J. Biol. Chem. 269:2440-2446(1994).
RN [2]
RP SEQUENCE OF 626-1895 AND 4196-5179 FROM N.A.
RC TISSUE=Colon;
RX MEDLINE=93016075; PubMed=1400449;
RA Gum J.R. Jr., Hicks J.W., Toribara N.W., Rothe E.-M., Lagace R.E.,
RA Kim Y.S.;
RT "The human MUC2 intestinal mucin has cysteine-rich subdomains located
RT both upstream and downstream of its central repetitive region.";
RL J. Biol. Chem. 267:21375-21383(1992).
RN [3]
RP SEQUENCE OF 1343-1895 AND 4176-4195 FROM N.A.
RX MEDLINE=91358717; PubMed=1885763;
RA Toribara N.W., Gum J.R. Jr., Culhane P.J., Lagace R.E., Hicks J.W.,
RA Petersen G.M., Kim Y.S.;
RT "MUC-2 human small intestinal mucin gene structure. Repeated arrays
RT and polymorphism.";
RL J. Clin. Invest. 88:1005-1013(1991).
CC -!- FUNCTION: COATS THE EPITHELIA OF THE INTESTINES, AIRWAYS, AND
CC OTHER MUCUS MEMBRANE-CONTAINING ORGANS. THOUGHT TO PROVIDE A
CC PROTECTIVE, LUBRICATING BARRIER AGAINST PARTICLES AND INFECTIOUS
CC AGENTS AT MUCOSAL SURFACES.
CC -!- SUBUNIT: MULTIMERIC.
CC -!- SUBCELLULAR LOCATION: SECRETED.
CC -!- TISSUE SPECIFICITY: COLON, SMALL INTESTINE, COLONIC TUMORS,
CC BRONCHUS, CERVIX AND GALL BLADDER.
CC -!- PTM: ALL CYSTEINE RESIDUES ARE INVOLVED IN INTRACHAIN OR
CC INTERCHAIN DISULFIDE BONDS (BY SIMILARITY).
CC -!- POLYMORPHISM: THE NUMBER OF REPEATS IS HIGHLY POLYMORPHIC AND
CC VARIES AMONG DIFFERENT ALLELES.
CC -!- SIMILARITY: THE N-TERMINAL DOMAIN SHOWS SOME SIMILARITY TO THAT
CC OF SILKWORM HEMOCYTIN.
CC -!- SIMILARITY: CONTAINS 2 VWFC DOMAINS.
CC -!- SIMILARITY: CONTAINS 1 C-TERMINAL CYSTINE KNOT-LIKE DOMAIN (CTCK).
CC ---

T11622
extensin class 1 precursor - cowpea
C;Species: Vigna unguiculata (cowpea)
C;Date: 16-Jul-1999 #sequence_revision 16-Jul-1999 #text_change 17-Nov-2000
C;Accession: T11622; S54155
R;Arsenijevic-Maksimovic, I.; Broughton, W.J.; Krause, A.
Mol. Plant Microbe Interact. 10, 95-101, 1997
A;Title: Rhizobia modulate root-hair-specific expression of extensin genes.
A;Reference number: Z17301; MUID:97155574
A;Accession: T11622
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-489 <ARS>
A;Cross-references: EMBL:X91836; NID:g1015936; PIDN:CAA62943.1; PID:g1015937
A;Experimental source: sub_species Red caloona
R;Arsenijevic-Maksimovic, I.; Broughton, W.J.; Krause, A.
submitted to the EMBL Data Library, April 1995
A;Description: A class of root-hair specific extensins involved in rhizobium/legume inte
A;Reference number: S54155
A;Accession: S54155
A;Status: preliminary
A;Molecule type: mRNA
A;Residues: 326-489 <AR2>
A;Cross-references: EMBL:X86030; NID:g791149; PID:g791150
C;Genetics:
A;Gene: Ext26G
C;Superfamily: hydroxyproline-rich glycoprotein
C;Keywords: glycoprotein; hydroxyproline
F;1-23/Domain: signal sequence #status predicted <SIG>
F;24-489/Product: extensin class 1 #status predicted <MAT>

Query Match 8.5%; Score 633; DB 2; Length 489;
Best Local Similarity 31.2%; Pred. No. 1.4e-19;
Matches 149; Conservative 33; Mismatches 252; Indels 44; Gaps 7;

```
Qy    354 PKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAP 413
          ||:  |    : |   | | |     |:|: |          | |   |  |   | | |
Db     39 PKQTPPYYYNAPPYYYKSPPP-----PSPSPP----------PPPYVHKYPPYYYKSPPP 83

Qy    414 TTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTP---KEPAPTAPKKPAPTT 470
           :|  | |   | | | :|  | |   | | | :  |:|  |   | | | :|   | |
Db     84 PSPSPPPPYVYKSPPPPSPSPPPPYYYKSPPPPS--PSPPPPYYYKSPPPPSPSPPPPYY 141

Qy    471 PKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSP 530
           | | | :|  | |   | | | :|  | |   || |              | :|  | |
Db    142 YKSPPPPSPSPPPPYVYKSPPPPSPSPPPPYYYKSPP-------------PPSPSPPPP 187

Qy    531 TTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPK 590
            | | | :|  | |   | | | :|  | |   | | | :   | |   | | | :|
Db    188 YYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYVYKSPPPPSPSPPPPYYYKSPPPPSPS 247

Qy    591 ETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAP 650
             |   |   | :|    |   : | | :|    |   : | | :|  | |   |:  |
Db    248 PPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPP 307

Qy    651 NTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELA 710
           :|  | |   | | | :|  | |   |   | :|    |   | | | :|  | |
Db    308 PSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYVYKSPPPPSPSPPPPYYY- 366

Qy    711 PTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTP 770
            | |   :   | |   |   | :|  | |   | | | :|    |   | | | :|
Db    367 ----KSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSP 422

Qy    771 KKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETPPP 828
            |     |  | |   :   | |   |   | :|  | |   | | | :|  |||
Db    423 SPP-----PPYVYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPP 475
```

Search completed: April 26, 2002, 16:18:41
Job time: 311 sec

```
Db   2258 -----STEQPTSTTPSGQSLTPMNSNSEVLTTSEPHVLSSSLSPDVSQSSTTPNNLSESS 2312

Qy    533 TKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTT------------------- 572
Db   2313 TVE----TPKTSSEVSLNSEEPSTTEAPTTLSPDILSTTTNNLSQSSTVSTEDRSEISSE 2368

Qy    573 -TKKPAPTAPK----------EPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTP 621
Db   2369 NSEKPT-SAPELVTSSVTHVASSSPDVPTES--SEPDDLTGSSTENIPEASSKQTISSTP 2425

Qy    622 EELAPTTPEEPTPTTPEEP-----------APTTPKAAAPNTPKEPAPTT---------- 660
Db   2426 TPDTTTASEEPTKSTSMSPDLSTTSNVLSESSTTPESSS----KSPVSSSTEGISVVTST 2481

Qy    661 --PKEPAPT---------TPKEPAP-----TTPKETAPTTPKGTAPTTLKEPAPTTPKKP 704
Db   2482 EFSKVPESTISSVLEEDLTKTTPSPILEETTTASETSEPLTEDSLTVSVRIHELTTSSEN 2541

Qy    705 APKELAPTTTKEPTSTTSDKPA---------PT-----------------TP--KGTAP- 735
Db   2542 VPKESESTTTSSESSKPSQEPAGILTSTVVVPTSSVSLITASEIEAITSNTPFKQGRTPI 2601

Qy    736 -TTPK-------EPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPT 787
Db   2602 TTSPKSLVKSTTSPSTVTSSEPSESTKRTTVSTTVSTTTPTEETTTSESLILTAAPSKPT 2661

Qy    788 -STTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIH 846
Db   2662 ESTTESSEAPTTPAKTSETKPSNVSSTSRKSTENVETSTSQSGSLESSTMSSTS------ 2715

Qy    847 KSPDESTP--ELSAEPTPKALE--NSPKEPGVPTTKTPAATK-----PEMTTTAKDKTTE 897
Db   2716 SEPETNAPAVTVSSEASSTTLEENSSTSSP----TSSEASVKLSSLFPESITSEAVTVSS 2771

Qy    898 RDLRTTPETTTAAPKMTKETATTTEKTTESKITATTT--------------QVTSTTTQ 942
Db   2772 R----APAEITMSSESHREISTVSSEPSEPEIPLSTTVSPNVVTASSIPSEEPILSSVTS 2827

Qy    943 DTTP-FKITT-----LKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPK 996
Db   2828 SSTPRVRLITGTPDDLIVSVTVPSHGNRRQNITASSV---PSNSTSPIILPSES-LTTPQ 2883

Qy    997 PQKPTKAPKKPTSTKKPKTMPRVRKPK---TTPTPRKMTSTMPELN-----PTSRIAEAM 1048
Db   2884 PPPTTTTTAKPATTSGKRGPPSIQPPAEMFTTPAP------PPPSNGGYGEETNQEEEQV 2937

Qy   1049 LQTTT 1053
Db   2938 TSTTT 2942
```

RESULT 14
I51618
nucleolar phosphoprotein - African clawed frog
C;Species: Xenopus laevis (African clawed frog)
C;Date: 13-Sep-1996 #sequence_revision 13-Sep-1996 #text_change 24-Sep-1999
C;Accession: I51618; S57757
R;Cairns, C.; McStay, B.
J. Cell Sci. 108, 3339-3347, 1995
A;Title: Identification and cDNA cloning of a Xenopus nucleolar phosphoprotein, xNopp180
A;Reference number: I51618; MUID:96019267
A;Accession: I51618
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: mRNA
A;Residues: 1-990 <CAI>
A;Cross-references: EMBL:X88927; NID:g895920; PIDN:CAA61368.1; PID:g895921
C;Genetics:
A;Gene: xNopp180
C;Superfamily: nucleolus-cytoplasm shuttle phosphoprotein
C;Keywords: phosphoprotein

Query Match 8.7%; Score 643.5; DB 2; Length 990;

Best Local Similarity 24.2%; Pred. No. 1.1e-19;
Matches 246; Conservative 148; Mismatches 424; Indels 199; Gaps 40;

```
Qy     88 SPPSSKKAPPPSG--------ASQTIKSTTKRSPKPPNKKKT------KKVIESEEITEE 133
Db     57 SPDAKKRKPPANGLPKKKSAKESSSSEDSSSEEDEPPAKKRAQPAGGKKPVVKAVQPKKA 116

Qy    134 HSVSENQESSSSSSSSSSSTIWK-IKSSKNSAANRELQKKLKVKDNKKNRT------KK 186
Db    117 KSSSEDSSDESDSEEETKKPPAKRPAQTPKVAAVKTPTQKKAKSSSSESSSSEDEASKKK 176

Qy    187 KPTPKPP---VVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPRPSLPPN 243
Db    177 QPVIKVPPKQAVVKAGLASNNG----KTADSSSSEDSDSPPAKKTAATK-------TPP- 224

Qy    244 SDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTS-- 301
Db    225 --TKPATAAKPQAKKTAGKKSSSREDSSDSSDEEQKTAKSKPKPDVYSAVP---PPTSVS 279

Qy    302 --KVLAKPTPKAETTTKGPALTTPKEPTPTTPKE--PASTTPKEPTPTTIKSAPTTPKEP 357
Db    280 KKKTLSQPGTKAKPESSDSSDSSDEEEQPAKKAKIVPAKAAASAPKPLAKKAETSTDSES 339

Qy    358 APTT-----------TKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKE-PAPTTTKSAP 405
Db    340 DSSSEDEKKSSVKLGVKAAP--KKAPA-------APDAKSTPVAAAKKSAPAKKASSSSD 390

Qy    406 TTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPA-PTTPKEPAPTAPK 464
Db    391 SDSSSNEETTTKPAAKTTPAKSA-ATPTSKTPTNGK-ATPTSKTPAKPGTPKTSTAKKDS 448

Qy    465 KPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTT 524
Db    449 SSSDSSDSSSDEETTTKPAAKTTPAKSAATPTSKTPTNSKATPTSKKTPAKPGTPKTSAA 508

Qy    525 PKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEP-APTTPKEPAPTTTKKPAPTAPKE 583
Db    509 KKDSSSSDSSDSSSDEKKTPA----KRAAKTTPAKPAAKTTPAKPA-------AKTTPAK 557

Qy    584 PAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 643
Db    558 PA----------------------AKSTPGKQVPTKKE----SSSSDSSDSSSEDEKKS 590

Qy    644 TPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKK 703
Db    591 SAKPAVKTT---PGKATSK---PVVASKPVP-AKKASSSSDSDSSEEETTKTTKPLTKLS 643

Qy    704 PAPKELAPTTTKEPTSTTSD-------KPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKG 756
Db    644 PAVKTLPPKKAESSSDSSSDSDSEKKTKPAKPPAKSATPVNTKAPAQNKASKASCSDSDS 703

Qy    757 TAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPK 816
Db    704 SSEEEGKSKQP-TGKSPAAKATAP-PKKNPVAVNKDKPSSSSSSDSSGDDEKQ----KPK 757

Qy    817 KPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPT 876
Db    758 QAA-------------------------AAKDVKQGAKAAKPTPKKAASSSSE---DS 787

Qy    877 TKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQV 936
Db    788 SSDEDVSKAKKTNTAVSKSPV----TTPKAVPAAKKESSSESSDSEDEKQGGKNTSTTKI 843

Qy    937 TSTTTQDTTPFKITTLKTTTLAPKVTTTK--KTITTTEIMNKPEETA-KPKDRATNS--- 990
Db    844 ANST-----------------PKAAAAECSEESSSSEDEGKANGTSGKRKRESTGNAEC 885

Qy    991 KATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTP-RKMTSTMPELNPTSRIAE 1046
Db    886 EAVTPE--------NKKLKAKSPNTFPKVNKKELKNTPFRRVVEEDIEINP--RMAD 932
```

RESULT 15

```
Qy    428 APTTPKEPTPTTPKEPAPTTKEPAPT--TPKEPAPTAPKKPAPTTPKEPAPTT-PKEPAP 484
Db   2153 ALSSPHPSSRTTESPPSPGTTTPGHTRGTSRTTATATPSKTRTSTLLPSSPTSAPITTVV 2212

Qy    485 TTTKEPS------------------------------------------------------ 491
Db   2213 TTGCEPQCAWSEWLDYSYPMPGPSGGDFDTYSNIRAAGGAVCEQPLGLECRAQAQPGVPL 2272

Qy    492 ------------------------------------------PTTPKEPAPTTTKSAP 507
Db   2273 RELGQVVECSLDFGLVCRNREQVGKFKMCFNYEIRVFCCNYGHCPSTPATSSTATPSSTP 2332

Qy    508 TTT--KEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTP 565
Db   2333 GTTWILTKLTTTATTTESTGSTATPSSTQGPPAGTPH--VSTTATTPTVTSSK----ATP 2386

Qy    566 KEPAPTTTKKPA-------PTAPKEPA---------------PTTPKETAPTTPKKLTP- 602
Db   2387 FSSPGTATALPALRSTATTPTATSFTAIPSSSLGTTWTRLSQTTTPMATMSTATPSSTPE 2446

Qy    603 --------------------TTPEKLAPTTPEKPAPTT-------PEELAPTTPEEP-- 632
Db   2447 TVHTSTVLTTTATTTGATGSVATPSSTPGTAHTTKVPTTTTTGFTVTPSSSPGTARTPPV 2506

Qy    633 ---TPTTPEEPAPTTPKAAAPNTPKEPAP-TTPKEPAPT----TPKEPAPT--TPKE--T 680
Db   2507 WISTTTTPTTSGSTVTPSSIPGTTHTPTVLTTTTQPVATGSMATPSSSTQTSGTPPSLIT 2566

Qy    681 APTTPKGTAPTTLKEPAPTTPKKPAPKELAPT-TTKEPTS----------TTSDKPAPTT 729
Db   2567 TATTITATGSTT--NPSSTPGTTPIPPELTTTATTPAATSSTVTPSSALGTTHTPPVPNT 2624

Qy    730 PKGT-----APTTP-------------------KEPAPTTPKEPAPTTPKGTAPT-TLK 763
Db   2625 TATTHGRSLSPSSPHTVRTAWTSATSGTLGTTHITEPSTGTSHTPAATTGTTTTSTPALS 2684

Qy    764 EPAP---TTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPTT--PKEP-------- 810
Db   2685 SPHPSSRTTESPPSP----GTTTPGHTTATSRTTATATPSKTRTSTLLPSQPTSAPITTV 2740

Qy    811 ------------------------------------------------------------ 810
Db   2741 VTTGCEPQCAWSEWLDYSYPMPGPSGGDFDTYSNIRAAGGAVCEQPLGLECRATAQPGVP 2800

Qy    811 ------------------------------------------APTTPKKPAPTTPETP 826
Db   2801 LGELGQVVECSLDFGLVCRNREQVGKFKMCFNYEIRVFCCNYGHCPSTPATSSTAMPSST 2860

Qy    827 PPTT---SEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTK----- 878
Db   2861 PGTTWILTELTTTATTTASTGSTATP-SSTP--GTAPPPKVLTSPATTPTATSSKATSSS 2917

Qy    879 -----------TPAATK-------------------PEMTTTAKDKTTERDLRTTPETT 907
Db   2918 SPRTATTLPVLTSTATKSTATSVTPIPSSTLGTTGTLPEQTTTPVATMSTIHPSSTPETT 2977

Qy    908 -TAAPKMTKETATTTEKT------------TESKITATTTQVTS-TTTQDTTP------ 946
Db   2978 HTSTVLTTKATTRATSSTSTPSSTPGTTWILTELTTAATTTAGTGPTATPSSTPGTTWIL 3037

Qy    947 FKITTLKTTTLAPKVTTT-KKTITTTEIMNKPEE----TAKPKDRAT--NSKATTPKPQ- 998
Db   3038 TELTTTATTTASTGSTATLSSTPGTTWILTEPSTTATVTAPPGSTATASSTQATAGTPHV 3097

Qy    999 -----KPTKAPKKPTSTKKPKT---MPRVRKPKTTPTPRKMT------------------ 1032
Db   3098 STTATTPTVTSSKATPSSSPGTATALPALRSTATTPTATSFTAIPSSSLGTTWTRLSQTT 3157

Qy   1033 -----------STMPELNPTSRIAEAMLQTT------TRPNQTPNSKLVEVNPKSEDAG- 1074
Db   3158 TPTATMSTATPSSTPETVHTSTVLTTTATTTGATGSVATPSSTPGTAHTTKVPTTTTTGF 3217

Qy   1075 -GAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSDETNICNGKPVDGLTTL 1133
```

```
                 :|    | | |::    |       |      |  : |      |              ||:
Db   3218 TATPSSSPGTALTPPVWISTTTTPTTTTP--TTSGSTVTPSSIPGTTHTARVLTTTTTTV 3275

Qy   1134 RNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGIPSPIDTVFTRCNCEGKT 1184
            |::               : || :   |:  | |  :  |  |     | |
Db   3276 ATGSM-------------ATPSSS---TQTSGTPPSLTTTATTITATGST 3309


RESULT  13
T34513
hypothetical protein ZK783.1 - Caenorhabditis elegans
C;Species: Caenorhabditis elegans
C;Date: 29-Oct-1999 #sequence_revision 29-Oct-1999 #text_change 29-Oct-1999
C;Accession: T34513
R;Favello, A.; Vaudin, M.
submitted to the EMBL Data Library, August 1994
A;Description: The sequence of C. elegans cosmid ZK783.
A;Reference number: Z21536
A;Accession: T34513
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-3507 <FAV>
A;Cross-references: EMBL:U13646; PIDN:AAC24418.1; GSPDB:GN00021; CESP:ZK783.1
A;Experimental source: strain Bristol N2; clone ZK783
C;Genetics:
A;Gene: CESP:ZK783.1
A;Map position: 3
A;Introns: 14/1; 48/2; 84/1; 196/3; 303/1; 381/1; 586/1; 605/1; 1175/3; 1207/1; 1409/
3504/1


  Query Match             8.9%;  Score 658.5;  DB 2;  Length 3507;
  Best Local Similarity   22.6%;  Pred. No. 8.7e-20;
  Matches  286;  Conservative  178;  Mismatches  484;  Indels  317;  Gaps   51;

Qy      3 LSSCAGRCGEGYSRD-ATC-------NCDYNCQHYMECCPDFKRVCTAELSCKGRCFESF 54
Db   1781 MGSCGCKCMAGYTGDGATCIKIEEEEPKSDKTA-----CTDEWSRLCELE---KKQCTVDE 1832

Qy     55 ERGREC-----------------DCDAQCKKYDKCCPDYESFCAEVHNPTSPPSSKKAPP 97
Db   1833 EEVPQCGACLPGHHPINGTCQSLQISGLCAQKNDCNKHAE--CIDIH-----PDSHFCSC 1885

Qy     98 PSG-------------------------ASQTIKSTTKRSPKPPNKKKTKKVIESEEIT 131
Db   1886 PDGFIGDGMICDDVDECNNAGMCDDENTKCENTIGSFNCVCLEGFKKVDEKCVVDEKKQP 1945

Qy    132 EEHSVSENQESSSSSSSSSSSTIWKIKSSKNSAANRELQKKLKVKDNKKNRTKKKPTPK 191
Db   1946 NREKIEIDEENSSSSNSG---------------------------------QEKPTTK 1970

Qy    192 PPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPRPSLPPN---SDTSK 248
Db   1971 GIVSSTS-----------ATSSESTTAEPHVTTSISSTTSTKDMTSSKSPENVTMSSESP 2019

Qy    249 ETSLTVNKETTV-ETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAP----TSKV 303
Db   2020 EVSTSSSKSTTASETTVSSTPSESSSSEAPLTSSPATTTEVITESSVKSTTPKEESSSEI 2079

Qy    304 LAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKS-APTTPKEPAPTTT 362
Db   2080 TVKLSSKSPEVTESSVKSSPSTPS-TTSQSVTSTVPETSKSTVLSSEAPVTSTSPTEVHT 2138

Qy    363 KSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPT 422
Db   2139 -SSETKPSLSASSTTGDTNSTTPSTSSLASVKSTSAPEGTSASVAPVKLSSLSPDVSQPS 2197

Qy    423 TPK-EPAPTTPKEPTPTTPKEPAPTTKEP-APTTPKEPAPTAPKKPAPTTPKEPAPTTPK 480
Db   2198 TKTFDATESSTVQASETSSGTSVKSTSEPESHVTKLSITSSNPSSSVPVTSPKSTPTVPE 2257

Qy    481 EPAPTTTKEPSPTTP--KEPAPTTTKSAPTTTKEPAPTTTKSAP------TTPKEPSPTT 532
```

```
  Query Match             9.0%;  Score 666.5;  DB 2;  Length 7962;
  Best Local Similarity   24.2%;  Pred. No. 9.1e-20;
  Matches  268;  Conservative  120;  Mismatches  491;  Indels  229;  Gaps   50;

Qy      74 PDYESFCAEVHNPTSPPSSKKAPPPSGASQTIKSTTKRS-------------PKPPNKK 119
           |: | |  |       |    | | |:   :  |  |::              |: | |
Db    6827 PEEEEFVPE--EEVLPEVKPKVPVPAPVPEIKKKVTEKKVVIPKKEEAPPAKVPEVPKKV 6884

Qy     120 KTKKVI--ESEEITEEHSVSENQESSSSSSSSSSSSTIWKIKSSKNSAANRELQKKLKVK 177
           : |::|  : ||:       | :|  |           :|     |    |     :|
Db    6885 EEKRIILPKEEEVLPVEVTEEPEEEPISEE---------EIPEEPPSIEEVEEVAPPRVP 6935

Qy     178 DNKKNRTKKKPTPKPPVVDEAGSGLDN--GDFKVTTP-DTSTTQHNKVSTSPKITTAKP- 233
           :  |    : ||| |  |:   : :     : ||  |          ||   ||    |
Db    6936 EVIKKAVPEAPTPVPKKVEAPPAKVSKKIPEEKVPVPVQKKEAPPAKVPEVPKKVPEKKV 6995

Qy     234 -INPRPSLPP---NSDTSKETSLTVNKETTVETK---ETTTTNKQTSTDGKE-------- 278
            :  : ::||    :   :: |:   :|  |: :   |       :   : :|
Db    6996 LVPKKEAVPPAKGRTVLEEKVSVAFRQEVVVKERLELEVVEAEVEEIPEEEEFHEVEEYF 7055

Qy     279 -------------KTTSAKETQSIEKT-------SAKDLAPTSKVLAKPTPKAETTTKG 317
                       :   :|  :||          |:::    |  | |         ||
Db    7056 EEGEFHEVEEFIKLEQHRVEEEHRVEKVHRVIEVFEAEEVEVFEKPKAPP--------KG 7107

Qy     318 PALT-----TPKEPTPTTP-KEPASTTPKEPTPTTIKSAPTTPKEP-APTTTKSAPTTPK 370
           | ::       | ||   | ||| :  |: |    ::     |:||  | |        ||
Db    7108 PEISEKIIPPKKPPTKVVPRKEPPAKVPEVPKKIVVEEKVRVPEEPRVPPTKVPEVLPPK 7167

Qy     371 EPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKK----PAPTTPKE 426
           |  |           | | |   | ||       |  |: |    |:|    | |  |:
Db    7168 EVVP---------EKKVPVPPAKKPEAP-----PPKVPEAPKEVVPEKKVPVPPPKKPEV 7213

Qy     427 PAPTTPKEPTPTTPKEPAPTTKEPAPTTP-------KEPAPTAPKKPAPTTPKEPAPTTP 479
           |    |: |    |::  |    | | :|       | :|:|| |  |  |  |    | |
Db    7214 PPTKVPEVPKAAVPEKKVPEAIPPKPESPPPEVFEEPEESPSAPPK-KPEVPPVRVPEVP 7272

Qy     480 KEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPT 539
           ||  |      :|  ||:|  |  |  |   ||  |       |  ||:|    ||
Db    7273 KEVVPEKKVPAAP--PKKPEVTPVK-VPEAPKEVVPEKKVPVP-PPKKPEVPPTK----- 7323

Qy     540 TPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKK 599
            |: |   |:|  |    |  |  |: | |   ::|   | :||    :|  |  | :
Db    7324 VPEVPKVAVPEKKVP----EAIPPKPESPPPEVFEEPEEVALEEPPAEVVEEPEPAAPPQ 7379

Qy     600 LT----PTTPEKLAPTT----PEKPAPTTPEELAPTTPEEPTP-TTPEEPAPTTPKAAAP 650
           :|       ||| ||      || |    ||       ||:  |   |::  |  | |  |
Db    7380 VTVPPKNPVPEKKAPAVVAKKPELPPVKVPEVPKEVVPEKKVPLVVPKK--PEAPPAKVP 7437

Qy     651 NTPKEPAP----TTPKEP------APTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTT 700
             |||  |      ||:|       |  ||:|     |  |  |    |   : | |
Db    7438 EVPKEVVPEKKVAVPKKPEVPPAKVPEVPKKPV----LEEKPAVP---VPERAESPPPEV 7490

Qy     701 PKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPT 760
           ::  |:|:||   :|  :   :|| |   :      |: | |  |:||    |:   |
Db    7491 YEE--PEEIAP---EEEIAPEEEKPVPVAEE----EEPEVPPPAVPEEPKKIIPEKKVP- 7540

Qy     761 TLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETA-------PTTPKEPAPT 813
           :|:|    ||:|  |:::      |          | |  |||             ||:  |
Db    7541 VIKKPEAPPPKEPEPEKV---IEKPKLKPRPPPPPPAPPKEDVKEKIFQLKAIPKKKVPE 7597

Qy     814 TPKKP--APTTPETPPPTTSEVS--TPTTTKEP-------TTIHKSPDESTPELSAEPTP 862
           |: |     ||   |   :|     |    ||        : : | |:|  |::      |
Db    7598 NPQVPEKVELTPLKVPGGEKKVRKLLPERKPEPKEEVVLKSVLRKRPEEEEPKVE----P 7653

Qy     863 KALENSPKEPGVPTTKTPAATKP----EMTTTAKDKTTERDLRTTPETTTAAPKMTKETA 918
           | ||   |:| ||     |   ||    |: | | :     :    ||   | |      |
Db    7654 KKLE-KVKKPAVP---EPPPPKPVEEVEVPTVTKRERKIPEPTKVPEIKPAIPLPAPEPK 7709

Qy     919 TTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNK-- 976
           |                          :: |:|   : |: |     :|   :  |
Db    7710 PKPEA----------------------EVKTIKPPPVEPEPTPIAAPVTVPVVGKKAE 7745

Qy     977 ----PEETAKPKDRATNSKATTPKPQKPTKAPKKPTS--TKKPKTMPRVRKPKTTPTP-- 1028
                 || ||||          || | :  :   :| |   | |   |   : |  |
Db    7746 AKAPKEEAAKPKGPIKGVPKKTPSPIEAERRKLRPGSGGEKPPDEAPFTYQLKAVPLKFV 7805

Qy    1029 ---RKMTSTMPELNPTSRIAEAMLQTTT 1053
              : :  |  |   :| | | ::  :|
Db    7806 KEIKDIILTESEFVGSSAIFECLVSPST 7833
```

RESULT 12
T45025
mucin MUC5B, tracheobronchial [imported] - human (fragment)
C;Species: Homo sapiens (man)
C;Date: 21-Jan-2000 #sequence_revision 21-Jan-2000 #text_change 21-Jul-2000
C;Accession: T45025
R;Desseyn, J.L.; Guyonnet-Duperat, V.; Porchet, N.; Aubert, J.P.; Laine, A.
J. Biol. Chem. 272, 3168-3178, 1997
A;Title: Human mucin gene MUC5B, the 10.7 kb large central exon encodes various alter
A;Reference number: Z22899; MUID:97166151
A;Accession: T45025
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-3570 <DES>
A;Cross-references: EMBL:Z72496; NID:g1834502; PIDN:CAA96577.1; PID:g1834503
A;Experimental source: placenta
C;Genetics:
A;Gene: MUC5B

```
  Query Match             8.9%;  Score 663;  DB 2;  Length 3570;
  Best Local Similarity   22.1%;  Pred. No. 5.8e-20;
  Matches  396;  Conservative  119;  Mismatches  578;  Indels  698;  Gaps   72;

Qy      30 MECCPDFKRVCTAELSCKGRCFESFERGRECDCDAQCKKYDKCCPDYE--SFCAEV-HNP 86
           :||  ||  ||               | ||      |  |: | | :||   ||    | |
Db    1581 VECSLDFGLVC--------------RNRE-----QVGKF-KMCFNYEIRVFCCNYGHCP 1619

Qy      87 TSPPSSKKAPPPS-------------------GASQTIKSTTKRSPKPPNKKKTKKVIE 126
           ::| :|  | | |                     |  | |:|  :  ||     ||:
Db    1620 STPATSSTATPSSTPGTTWILTEQTTAATTTATTGSTAIPSSTPGTAPPP------KVLT 1673

Qy     127 SEEITEEHSVSENQESSS-----------SSSSSSSSSTIWKIKSS----KNSAANR--- 168
           |:  |   : |:   |||           |::: |::::   | ||      :: ||
Db    1674 SQATTPTATSSKATSSSSPRTATTLPVLTSTATKSTATSFTPIPSSTLGTTGTSQNRPPH 1733

Qy     169 -----------------ELQKKLKVKDNKKNRTKKKPTPKPP-----VVDE------AGS 200
                           |  |      |    ||        :: |       :
Db    1734 PMATMSTIHPSSTPETTHTSTVLTTKATTTRATSSMSTPSSTPGTTWILTELTTAATTTA 1793

Qy     201 GLDNG-----------------DFKVTTPDTSTTQHN---------KVSTSPKITTAKPI 234
           | :|                  || |  ||   :          || ||   ||   |
Db    1794 ALPHGTPSSTPGTTWILTEPSTTATVTVPTGSTATASSTRATAGTLKVLTS-TATTPTVI 1852

Qy     235 NPR-------------PSLPPNSDTSKETSLTVNKETTVET-----KETTTTNKQTSTDG 276
           : |             |:|    : |   ||:|    ::: |       :|||     ||
Db    1853 SSRATPSSSPGTATALPALRSTATTPTATSVTAIPSSSLGTAWTRLSQTTTPTATMSTAT 1912

Qy     277 KEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTK----------------GPA 319
           |      | ::  |:|      |      ||   |||                  | |
Db    1913 PSSTPETVHTSTVLTTTATTTRTGSVATPSSTPGTAHTTKVPTTTTTGFTATPSSSPGTA 1972

Qy     320 LTTP----KEPTPT------TPKE-----------------------PASTTPKEPTP 344
           || |       |||      ||                        |:|:|    ||
Db    1973 LTPPVWISTTTPTTRGSTVTPSSIPGTTHTATVLTTTTTTVATGSMATPSSSTQTSGTP 2032

Qy     345 TTIKSAPTTPKEPAPTTTKSAP--TTPKEPAPTTT-KEPAPT----TPKEPAPTTTKEPA 397
           :: :  ||       ||  |:   |||  |  |||   || |    ||      ||   |
Db    2033 PSLTTTATTITATGSTTNPSSTPGTTPIPPVLTTTATTPAATSSTVTPSSALGTTHTPPV 2092

Qy     398 PTTTKSA------PTTP--------------------KEPAPTTPKKPAPTT----PKEP 427
           | || :        |::|                   ||:  |   || ||    |  |
Db    2093 PNTTATTHGRSLPPSSPHTVPTAWTSATSGILGTTHITEPSTGTSHTPAATTGTTQPSTP 2152
```

```
Biochem. Biophys. Res. Commun. 165, 644-649, 1989
A;Title: Sequence analysis of the 5' region of the human DF3 breast carcinoma-associated
A;Reference number: A36735; MUID:90088473
A;Accession: A36735
A;Molecule type: mRNA
A;Residues: 1-142,'Q',144-162,'Q',164-168 <ABE>
A;Cross-references: EMBL:M31823; NID:g181542; PIDN:AAA35757.1; PID:g181543
R;Masuzawa, Y.; Miyauchi, T.; Hamanoue, M.; Ando, S.; Yoshida, J.; Takao, S.; Shimazu, H
J. Biochem. 112, 609-615, 1992
A;Title: A novel core protein as well as polymorphic epithelial mucin carry peanut agglu
A;Reference number: JX0235; MUID:93123189
A;Accession: PX0066
A;Molecule type: mRNA
A;Residues: 998-1011,'ES',1014-1017;1018-1032,'T',1034-1037;1038-1057 <MAS>
A;Experimental source: gastric carcinoma cell
R;Zrihan-Licht, S.; Baruch, A.; Elroy-Stein, O.; Keydar, I.; Wreschner, D.H.
FEBS Lett. 356, 130-136, 1994
A;Title: Tyrosine phosphorylation of the MUC1 breast cancer membrane proteins cytokine r
A;Reference number: S51026; MUID:95080414
A;Contents: annotation
A;Note: undetermined tyrosine residues in the carboxyl-terminal non-repetitive region ar
C;Comment: This protein is length polymorphic. Individuals may have between 21 and 125 c
 partial repeats. The repeat shown is defined by SmaI nuclease sites.
C;Comment: Serine and threonine residues in the tandem repeat domain are extensively gly
C;Comment: For an alternative splice form without a tandem repeat domain, see PIR:S48146
C;Genetics:
A;Gene: GDB:MUC1; PUM
A;Cross-references: GDB:120705; OMIM:158340
A;Map position: 1q21-1q23
A;Introns: 20/1; 62/3; 1165/3; 1184/2; 1230/1; 1270/3; 1320/3
C;Superfamily: polymorphic epithelial mucin
C;Keywords: alternative splicing; duplication; glycoprotein; phosphoprotein; polymorphis
F;1-1344/Product: mucin 1 precursor, splice form A #status predicted <PREA>
F;1-62/Region: mucin 1 amino-terminal non-repetitive
F;1-23/Domain: signal sequence #link PREA #status predicted <SIGA>
F;1-19,29-32/Domain: signal sequence #link PREB #status predicted <SIGB>
F;1-19,29-1344/Product: mucin 1 precursor, splice form B #status predicted <PREB>
F;1-19,29-212,1033-1344/Product: mucin 1 precursor, epithelial tumor antigen splice form
F;138-1017/Region: 20-residue repeats (GSTAPPAHGVTSAPDTRPAP)
F;1143-1344/Region: mucin 1 carboxyl-terminal non-repetitive
F;1245-1272/Domain: transmembrane #status predicted <TRM>
F;1046,1064,1118,1144,1222/Binding site: carbohydrate (Asn) (covalent) #status predicted
F;1213/Binding site: phosphate (Tyr) (covalent) #status predicted

  Query Match             9.1%;  Score 672;  DB 1;  Length 1344;
  Best Local Similarity   27.8%;  Pred. No. 9.6e-21;
  Matches  310;  Conservative   78;  Mismatches  472;  Indels  256;  Gaps   56;

Qy    252 LTVNKETTVETKETTTTNK--QTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 308
Db     15 LTVLTATTAPKPATVVTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 74

Qy    309 P-KAETTTKGP--ALTTPKEP--------------TPTTPKEPASTTP----------KE 341
Db     75 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVTSAPDNK 134

Qy    342 PTPTT-------IKSAPTTPKEPAPTTT-------KSAPTTPKEPAPTTTKEP------A 381
Db    135 PAPGSTAPPAHGVTSAPDT--RPAPGSTAPPAHGVTSAPDT--RPAPGSTAPPAHGVTSA 190

Qy    382 PTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTP----------KKPAPTTPKEP---- 427
Db    191 PDT--RPAPGSTAPPAHGVT-SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGV 247

Qy    428 --APTTPKEPTPTTP----------KEPAPTTKEP-------APTTPKEPAPTAPK---- 464
Db    248 TSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGV 307

Qy    465 KPAPTTPKEPAPTTPK----EPAPTTTKEPSPTTPKEPAPTT---TKSAPTTTKEPAPTT 517
Db    308 TSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGV 367

Qy    518 TKSAPTTPKEPSPTTTKEP------APTTPKEPAPTTPK----KPAPTTPKEPAPTTPK- 566
Db    368 T-SAPDT--RPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPA 424

Qy    567 ---EPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPE- 622
Db    425 HGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPA 484

Qy    623 ---ELAPTTPEEPTPTTP----------EEPAP--TTPKA----AAPNTPKEPAPTTPK- 662
Db    485 HGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPA 544

Qy    663 ---EPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTP----------KKPAPKEL 709
Db    545 HGVTSAPDTRPAPGSTAP----PAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGST 600

Qy    710 APTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTPKEPAPTTPKG----TAPTTLK 763
Db    601 AP-PAHGVTSAPDTRPAPGSTAPPAHGVTS----APDTRPAPGSTAPPAHGVTSAPDTRP 655

Qy    764 EPAPTTP----------KKPAPKELAPTTTKGPTSTTSDKPAP--TTPK----ETAPTTP 807
Db    656 APGSTAPPAHGVTSAPDTRPAPGSTAP-PAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTR 714

Qy    808 KEPAPTTP----------KKPAPTTPETPPPTTSEVSTPTTTKEPTT-------IHKSPD 850
Db    715 PAPGSTAPPAHGVTSAPDTRPAPGS--TAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPD 772

Qy    851 ESTPELSAEPTPKALENSPKEPGVPTTKTPAA--------TKPEMTTTAKDK---TTERD 899
Db    773 TRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPD 832

Qy    900 LRTTPETT-------TAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTP--FKIT 950
Db    833 TRPAPGSTAPPAHGVTSAP----DTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT 888

Qy    951 TLKTTTLAPKVTTTKKTITTT--EIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPT 1008
Db    889 SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAP--GSTAPPAHG 946

Qy   1009 STKKPKTMPRVRKPKTTPTPRKMTSTMPELNP----TSRIAEAMLQT-TTRP---NQTPN 1060
Db    947 VTSAPDTRP---APGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPP 1003

Qy   1061 SKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSDETN 1120
Db   1004 AHGVTSAPDTRPAPGSTAPPAH----------GVTSAPDNRPALGSTA----PPVHNVTS 1049

Qy   1121 ICNGKPVDGLTTLRNGTLVAFRGHYFWMLSPFSPPS 1156
Db   1050 ASGSASGSASTLVHNGTSARATTTPASKSTPFSIPS 1085


RESULT  11
I38346
elastic titin - human (fragment)
C;Species: Homo sapiens (man)
C;Date: 29-May-1998 #sequence_revision 29-May-1998 #text_change 21-Jul-2000
C;Accession: I38346
R;Labeit, S.; Kolmerer, B.
Science 270, 293-296, 1995
A;Title: Titins: giant proteins in charge of muscle ultrastructure and elasticity.
A;Reference number: A57430; MUID:96026330
A;Accession: I38346
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: mRNA
A;Residues: 1-7962 <RES>
A;Cross-references: EMBL:X90569; NID:g1017426; PIDN:CAA62189.1; PID:g1017427
C;Genetics:
A;Gene: GDB:TTN
A;Cross-references: GDB:127867; OMIM:188840
A;Map position: 2q31-2q31
```

```
Qy      49 RCFESFERGRECDCDAQCKKYDKCCPDYESFCAEVHNPTSPPSSKKAPPPSGASQTIKST 108
Db      34 RDFKIVKHIKKNTCTCSC----KCVPD------------APSN-----PFDVSTTISS- 70

Qy     109 TKRSPKPPNKKKTKKVIESEEITEEHSVSENQESSSSSSSSSSSSTIWKIKSSKNSAANR 168
Db      71 ----------------INNDNVDIGPSGDSNPTGSSWFQEIEATVGGQTVKSEHNIDSSV 114

Qy     169 ELQKKLKVKDNKKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQ---HNKVSTS 225
Db     115 EVEKKV---------------------------------TTSTDASTTNAPTTGKDSTT 140

Qy     226 PKITTA-KPINPRPSLPPNSDT-------SKETSLTVNKETTVETKETTTTNKQTSTDGK 277
Db     141 PEIITGIVVINSKSESVTDMSTTRFSTTLSPTTELLTSPETLVSTDSSTST-EQTSPDNT 199

Qy     278 EKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPAST 337
Db     200 TEIASPMETNTTTEATTTSVEPSVSTLAS---EDETTVTAIAEST------TTVIAEVST 250

Qy     338 TPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPA 397
Db     251 TTEEPTTT-----------AESTTKKSTT----KAPATTEEPTPTTTEE--VTTTEAET 292

Qy     398 PTTTKSAPTTPKEPAPTTP----KKPAPTTPKEPAPTTPKEP-TPTTPKEPAPTTKEPAP 452
Db     293 STTTSSETSTEK---PTTPLIDNKIAGPATGK---PETTHFPVTGTTPN--FDTATETPF 344

Qy     453 TTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKE-----PSPTTPKEPAPTTTKSAP 507
Db     345 VAKSEDKMTLSKTAATETTQQTTEVT-DGPEKETTKNVSIEIPITTVPLVETTSTTSTAS 403

Qy     508 -------TTTKEPAPT----TTKSAPTTPKEPSPTTTKE---PAPT-------TPK---- 542
Db     404 KESDGFHTTLKLKVTTADSDSTESATTVKPFNEETTTKSHVVPKPTKKGTVKVTPKLELS 463

Qy     543 --EPAPTTPKKPAP--------------------TTPKE--------------PAPTTP 565
Db     464 FDEPTEIT-KAPHPGKLLEKKTYHFVLSDNFARYSEAKENDDYNHLDYNHYREAKEPTTT 522

Qy     566 KEPAP----TTTKKPAPTA--PKEPAPTT---PKETAPTTPKKLTPTTPEKLA---PTTP 613
Db     523 EESSTTEEVTTTEEPANTGNPPTTENPTTTEQPTSTAESTTTALPFTTEQTVTTEEPTTA 582

Qy     614 EKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPA 673
Db     583 EKSTATQ----KPTTTQESVST---EKTSTTKKA---STTEE--PTTTDEPTTTT---ES 627

Qy     674 PTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAP----TT 729
Db     628 STTGKATTPELSTTSEETTTTELKITTE--------GSTTTEEPTTTAIFAEASTGIITT 679

Qy     730 PKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTP------------------- 770
Db     680 DEETTSTTSTTPEITSTKE--IVTESAITQTSVSVVESSTPRQLPERWKAIVNKFKHNLE 737

Qy     771 ----KKPAPKELAPTTTKGP----------------TSTTSDKPAPTTPKETAPTTPKEP 810
Db     738 VLKEKKRLLKEKESTSTTGSDSSETTTVVAENIDEVTTTEKEKVVQTTPITTEKSTTQEE 797

Qy     811 APTTPKKPAPTTPE--TPPPTTSE-VSTPTTTKEPTTIHKSPDESTPELSAEPTPKALEN 867
Db     798 TTTTTTTTEKTTSKTTTEKPTTSESATTETTTSEPST-----TEST-------------- 838

Qy     868 SPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTA----------------- 909
Db     839 --------TVDTSSATTEESSTAAETTTTSAE---TSETTTSESAAFITGESPENTALQS 887

Qy     910 --------------------APKMTKETATTTEKTTESKITATTTQ-VTST---TTQDT 944
Db     888 SSQKSEENESSAEKPGARRDFVPKKHKTTVKPAETTSAVAASTTTTEPITTTEKSTTLET 947

Qy     945 TPFKITTLKTTTLAPKVT---TTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ--- 998

                                                                          
Db     948 TPIEATTLNEVTGPAFVTGAPVDETTINTLELLSK----------INNTQISQPKPTDIS 997

Qy     999 -------------KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELN----- 1039
Db     998 KTDALSSLISGLIGSFTKAPMAPTI------------HTTTDAAFVTATEASLNDGSDK 1044

Qy    1040 -------PTSRIAEA 1047
Db    1045 KIIDEAQPTDEIRRA 1059
```

RESULT 10
A35175
mucin 1 precursor, repetitive splice form A [validated] - human
N;Alternate names: breast carcinoma-associated DF3 antigen; core protein KP39; episia
ncreatic mucin; polymorphic epithelial mucin (PEM)
N;Contains: mucin 1 precursor, epithelial tumor antigen splice form; mucin 1 precurso
C;Species: Homo sapiens (man)
C;Date: 20-Apr-2000 #sequence_revision 20-Apr-2000 #text_change 02-Jun-2000
C;Accession: A35175; B35175; A35886; A35887; S10572; S40293; A36735; PX00066; S10218;
R;Ligtenberg, M.J.L.; Vos, H.L.; Gennissen, A.M.C.; Hilkens, J.
J. Biol. Chem. 265, 5573-5578, 1990
A;Title: Episialin, a carcinoma-associated mucin, is generated by a polymorphic gene
A;Reference number: A35175; MUID:90202794
A;Accession: A35175
A;Molecule type: mRNA
A;Residues: 1-952,1033-1344 <LIG1>
A;Cross-references: GB:M32738; GB:J05288; NID:g182121; PIDN:AAA35804.1; PID:g182124;
A;Experimental source: splice form A
A;Note: GenBank entries HUMEPISIA1 and HUMEPISIA2 present only the amino-and carboxyl
A;Accession: B35175
A;Molecule type: mRNA
A;Residues: 1-19,29-952,1033-1344 <LIG2>
A;Cross-references: GB:M32739; GB:J05288; NID:g182126; PIDN:AAA35806.1; PID:g182129;
A;Experimental source: splice form B
A;Note: GenBank entries HUMEPISIB1 and HUMEPISIB2 present only the amino-and carboxyl
R;Gendler, S.J.; Lancaster, C.A.; Taylor-Papadimitriou, J.; Duhig, T.; Peat, N.; Burc
J. Biol. Chem. 265, 15286-15293, 1990
A;Title: Molecular cloning and expression of human tumor-associated polymorphic epith
A;Reference number: A35886; MUID:90368715
A;Accession: A35886
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 1-19,29-992,1033-1344 <GEN>
A;Cross-references: GB:J05581; NID:g188869; PIDN:AAA59876.1; PID:g188870
A;Note: GenBank entry HUMMUCAB includes one copy of the tandemly repeated sequence
R;Lan, M.S.; Batra, S.K.; Qi, W.N.; Metzgar, R.S.; Hollingsworth, M.A.
J. Biol. Chem. 265, 15294-15299, 1990
A;Title: Cloning and sequencing of a human pancreatic tumor mucin cDNA.
A;Reference number: A35887; MUID:90368716
A;Accession: A35887
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 1-19,29-1109,'S',1111-1339,'A',1341-1344 <LAN>
A;Cross-references: GB:J05582; NID:g189598; PIDN:AAA60019.1; PID:g189599
A;Note: GenBank entry HUMPANMU contains four fewer copies of the tandemly repeated se
R;Wreschner, D.H.; Hareuveni, M.; Tsarfaty, I.; Smorodinsky, N.; Horev, J.; Zaretsky,
Eur. J. Biochem. 189, 463-473, 1990
A;Title: Human epithelial tumor antigen cDNA sequences. Differential splicing may gen
A;Reference number: S10571; MUID:90276413
A;Accession: S10572
A;Molecule type: mRNA
A;Residues: 1-19,29-155,'P',157-175,'P',177-182,'A',184-212,1033-1344 <WRE>
A;Cross-references: EMBL:X52229; NID:g37053
R;Wreschner, D.H.
submitted to the EMBL Data Library, March 1990
A;Reference number: S40293
A;Accession: S40293
A;Molecule type: mRNA
A;Residues: 1-19,29-155,'P',157-175,'P',177-182,'A',184-212,1033-1037,'A',1039-1344 <
A;Cross-references: EMBL:X52229; NID:g37053; PIDN:CAA36478.1; PID:g37054
R;Abe, M.; Siddiqui, J.; Kufe, D.
```
```

```
Qy    739 -KEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPT 797
Db   1001 MSSPPPPEVKSPPPPAPVSSPPPPVKSPPPPAPVSSPP---PPVKSPPPPAPVSSPPPPV 1057

Qy    798 -TPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPEL 856
Db   1058 KSPPPPAPISSPPPPVKSPPPPAPVS--SPPPPVKSPPPPAPVSSPPPPIKSPPPPAPVS 1115

Qy    857 SAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPET 906
Db   1116 SPPPAP------VKPPSLP-PPAPVSSPPPVVTPAPPKKEEQSLPPPAES 1158
```

RESULT 8
T18535
high molecular mass nuclear antigen - chicken (fragment)
C;Species: Gallus gallus (chicken)
C;Date: 15-Oct-1999 #sequence_revision 15-Oct-1999 #text_change 15-Oct-1999
C;Accession: T18535
R;Shimada, K.; Harata, M.; Mizuno, S.
J. Cell Sci. 110, 3031-3041, 1997
A;Title: A nuclear matrix-associated high molecular mass nuclear antigen, HMNA, of chick
A;Reference number: Z18955; MUID:9803440
A;Accession: T18535
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: mRNA
A;Residues: 1-1151 <SHI>
A;Cross-references: EMBL:D88440; NID:d1177138; PID:d1025045; PIDN:BAA24137.1

Query Match 9.4%; Score 695; DB 2; Length 1151;
Best Local Similarity 25.3%; Pred. No. 9.4e-22;
Matches 304; Conservative 108; Mismatches 466; Indels 324; Gaps 59;

```
Qy     81 AEVHNPTSPPSSKK------APPPSGASQTIKSTTKRSPKPPNKKKTKKVIESEEITEEH 134
Db     39 AELH-PAAPQPPPKWVPIGGAPPPPGTEPT-------PPSKPTDGADAAPKASAELTSPP 90

Qy    135 SVSENQESSSSSSSSSSSSTIWKIKSSKNSAANRELQKKLKVKDNKKNRTKKKPTPKPPV 194
Db     91 PASPSPPDGPKAPSGAGEAEAGTPPPSQGPAG----------------------TPPP-- 126

Qy    195 VDEAGSGLDNGDFKVTTPDTSTTQHNKVSTS--PKITTAKPINPRPSLPPNSDTSKETSL 252
Db    127 -SQGAAGAPKGDGTAQPSGTKSGADGKPAAQDVPKATTA-ATEARP--------ASAASP 176

Qy    253 TVNKETTVETKETTTTNK--QTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPK 310
Db    177 TVPKATAEATAVTAASQSAPKAATDAAAVTAA---SQSAPKATV-EVKPAAAAVAKEAKA 232

Qy    311 AETTTKGPALTTPKEPTP-TTPKEPASTTPKEP----TPTTIKSAPTTPKEPAPTT---- 361
Db    233 VTAAAAAPKATAEAKPAPVTSPTIPCSSAEAKPLTAASPTASKA--TAEAKPVPATASLM 290

Qy    362 -------TKSAPTTPKEPAPTTTK----------------------EPAPTTPKEPAPT 391
Db    291 ATKVTAEAKPAPSPSVPKATTDTKAVTATAPKAGPDVKPAVAVCAEAKPAPPPPPQQLPK 350

Qy    392 TTKEPAPTTTKSAPTT------PKEPA-------PTTPKKPAPTTPKEPAPTTPKEPTPT 438
Db    351 AAAAAAPTGTELKPATAPPHGSPRANSHTVTVTPPNVPRAAAATVP--TAGAVPKASTGT 408

Qy    439 T----PKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEP---APTTTKEPS 491
Db    409 TPAAAPQQPVP---KAAPVTPPSPQQAVPRAATAAA----APVTPQQPVTKAATTTNATP 461

Qy    492 PTTPKEPAPTTTKSAPTTTKEPAP---TTTKSAPTTPKEPSPTTTKEP----APTTPKEP 544
Db    462 PPQPIPKAATTTTATPVTPQQPIPKAGTDAAPPPAVPKAPSDGRAATPGVPNAATDPQKP 521

Qy    545 APTTPKKPAPTTPKEPAP---------TTPKEPAPT-TTKKPAPTAPK-------EPAPT 587
```

```
Db    522 PPTPQSVPSAVTEPKPQPRAAPPPSNEATPAVPSPSPNLKSPLPTIPKPVPLMALTPQPV 581

Qy    588 TPK---ETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELA-------PTTP-------- 629
Db    582 TAQMVTQLAATKPSPIVPKASPK-ALMTPPPPPGLPRALAAAKLLGLPSSPVASAMHAK 640

Qy    630 --EEPTPTTP----EEPAPTTP---------KAAAPNTPKE----------PAPTTPKEP 664
Db    641 VTPRPLPASPVPMAASPASLGPDAARVALATNAASPGAKPEAAGGNGTLMAPMGAANTQM 700

Qy    665 APTTPKEPAPTTPKETAPT------------TPKGTAPT------------------- 691
Db    701 APIGAAGAAQTAPMGAAHTHVSPMGAGGATQMSPTGAANTHMSPIGAGGATQMSPMGAAN 760

Qy    692 TLKEPAPTTPKKPAPKELAPTTTKEP--------TSTTSDKPAPTTPKG--TAPTTPKEP 741
Db    761 TQMSPMGATTTQMSPMGAAATTQPSPMGAAATQVTATSAGNTMQVSPMGAATPPQTPSVG 820

Qy    742 APTTPKEPAPTTPKGTAPTTLKEP--APTTPKKPAPKELAPTTTKGP----TSTTSDKP- 794
Db    821 AATTP-QPSPM----GAATTLMSPMGAATTPQ---PSPMGAVTTQPPPMAATNTTQPPPM 872

Qy    795 APTTPKET---APTTPKEP---APTTPKKP-APTTPETPP-------PTTSEVSTPTTTK 840
Db    873 AASTPQSTPMGAATTTQSPPMGATTTQSPPMGASTPQAPPTVAGSPTPPPPIPPSPTAQT 932

Qy    841 EPTTIHKSPDESTPEL--SAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTER 898
Db    933 SPQPMSKSPPPDPPKAPSAAAQTSPAAHVANASPGV-TAVSPA---PIGVTEASPSADGA 988

Qy    899 DLRTTPETTTAAPKMT-KETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTL 957
Db    989 RLSPGPTAATDGPKASPAATADVTEAATD--VTAAATAVPA----EAAP---------TK 1033

Qy    958 APKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMP 1017
Db   1034 AKRSSSSSSSSSSSSSSSSSSSSSSSDSDSSSSSSESNPASPA---------------P 1078

Qy   1018 RVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAE 1077
Db   1079 AVGDGQQQMTPGAAQSVPP-------VTEAAVQ--------------EAAAAAAAAAGAE 1117

Qy   1078 GE 1079
Db   1118 RE 1119
```

RESULT 9
T25697
hypothetical protein F16F9.2 - Caenorhabditis elegans
C;Species: Caenorhabditis elegans
C;Date: 15-Oct-1999 #sequence_revision 15-Oct-1999 #text_change 18-Feb-2000
C;Accession: T25697
R;Fulton, B.
submitted to the EMBL Data Library, August 1996
A;Description: The sequence of C. elegans cosmid F16F9.
A;Reference number: Z20071
A;Accession: T25697
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-1229 <FUL>
A;Cross-references: EMBL:U67956; PIDN:AAB07691.1; GSPDB:GN00028; CESP:F16F9.2
A;Experimental source: strain Bristol N2; clone F16F9
C;Genetics:
A;Gene: CESP:F16F9.2
A;Map position: X
A;Introns: 35/1; 361/1; 384/1; 482/2; 517/1; 971/1; 1021/2; 1179/3

Query Match 9.1%; Score 676.5; DB 2; Length 1229;
Best Local Similarity 27.2%; Pred. No. 5.7e-21;
Matches 330; Conservative 95; Mismatches 385; Indels 405; Gaps 56;

```
  Query Match             10.9%;  Score 809;  DB 1;  Length 1367;
  Best Local Similarity   27.5%;  Pred. No. 2.3e-26;
  Matches  328;  Conservative  112;  Mismatches  526;  Indels  226;  Gaps   53;

Qy      61 DCDAQCKKYDK---CCPDYESFCAEVHNPTSPPSSKKAPPPSGASQTIKSTTKRSPKPPN 117
           |    |  ||       |:  |   :    :  :| :   |  |::  |||
Db     174 DLSTGCNNYDNQGHSQTDFPGFYWNIDCDNNCGGTKSSTTTSSTSES--STT-------- 223

Qy     118 KKKTKKVIESEEITEEHSVSENQESSSSSSSSSSSSTIWKIKSSKNSAANRELQKKLKVK 177
             |    ||   |   | |    ||:| ||:|||:|      :  :       |:
Db     224 ---TSSTSESSTTTSSTSESSTTTSSTSESSTSSSTT------APATPTTTSCTKEKPTP 274

Qy     178 DNKKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPR 237
           : ||:|||  ||  |              ||| |      | :|: |  | |   |
Db     275 PTTTSCTKEKPT--PPHHD-----------TTPCT-----KKKTTTSKTCTKKTTTPV 314

Qy     238 PSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDL 297
           |:  |:| |:: :|       | |  ::||          :::||  | |  ::|:
Db     315 PT--PSSSTTESSS------APVPTPSSSTT----------ESSSAPVTSSTTESSS--- 353

Qy     298 APTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEP 357
           ||       ||| : ||    |  |      ::    :|||      |    |: ||
Db     354 APV------PTPSSSTTESSSAPVTSSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSS 407

Qy     358 APTT-----TKSAP----TTPKEPAPTTTKEPAPTTPKEPAP----TTTKEPAPTTTKSA 404
           || |     : |||     ||    || |:     ||    ||    ||    ||  | |:
Db     408 APVTSSTTESSSAPVTSSTTESSSAPVTSS----TTESSSAPVTSSTTESSSAPVPTPSS 463

Qy     405 PTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTP---TTPKEPAPTTKEPAPTTPKEPAPT 461
           ||   || |   : ||    ||     |||   ||    || |   : ||    ||
Db     464 STTESSSAPVT----SSTTESSSAPV----PTPSSSTTESSSAPVT---SSTTESSSAPV 512

Qy     462 APKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSA 521
                 ||      :  ||||  :   | ||    || |:    ||    ||  | |:
Db     513 -------PTPSSSTTESSSAPAPTPS---SSTTESSSAPVTSS---TTESSSAPVPTPSS 559

Qy     522 PTTPKEPSPTTTKEPAPTTPKEPAPT-TPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTA 580
           ||    :| |:     ||    ||  ||       :   | ||    |: :|||:  :  |
Db     560 STTESSSTPVTSS----TTESSSAPVPTPSSSTTESSSAPVPT----PSSSTTESSSAPA 611

Qy     581 PKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPT-TPEEPTPTTPEE 639
           |   : ||   :|| |    : ||    || |   : ||    ||  ||   | :
Db     612 PTPSSSTTESSSAPVT----SSTTESSSAP-VPTPSSSTTESSSAPVPTPSSSTTESSSA 666

Qy     640 PAP-----TTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPT-TPKGTAPTTL 693
           | |     ||  ::|| |    : ||    || |    : ||   :||  ||  :   :
Db     667 PVPTPSSSTTESSSAPVT----SSTTESSSAPVT----SSTTESSSAPVPTPSSSTTESS 718

Qy     694 KEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAP---TTPKEPA 750
             | ||          || |   :|||    || |   |  ::   | |   ||     |
Db     719 SAPVPTPSSSTTESSSAPVPTPS-SSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSSA 777

Qy     751 PT-TPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKE 809
           |  ||  :   :   | ||          :||  |   :|  :      :||  ::  :
Db     778 PVPTPSSSTTESSSAPVPTPSSSTTESSVAPVPTPSSSSNITSSAPSSTPFSSSTESSSV 837

Qy     810 PAP-----TTPKKPAPTTPETPPPTTSEVSTPT-----TTKEPTTIHKSPDESTPELSAE 859
           . | |     ||    || :  |   : : | ||:      |:  |::|   |  || |  :
Db     838 PVPTPSSSTTESSSAPVSSSTTESSVAPVPTPSSSSNITSSAPSSI---PFSSTTESFST 894

Qy     860 PTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETAT 919
           |      :| | ||  |  : ::|       |  |:      ||  |||      |
Db     895 GT-TVTPSSSKYPGSQTETSVSSTTETTIVPTKTTTSVTTPSTTTITTTVCSTGTNSAGE 953

Qy     920 TTEKTTESKITAT--TTQVTSTTTQDTTPFKITTLKTTTLAPKVTT---TKKTITTT-EI 973
           ||   :  :| |  ||  || ||  ||     |  | | :   ||   : ||||||
Db     954 TTSGCSPKTVTTTVPTTTTTSVTTSSTTTITTTVCSTGTNSAGETTSGCSPKTITTTVPC 1013

Qy     974 MNKPEETAKPKDRATNSKATTPKPQKP----------------TKAPKKPTSTKKPKTM 1016
             | |||        |::||  |  |                 ||   : |:|   | :
Db    1014 STSPSETA--------SESTTTSPTTPVTTVVSTTVVTTEYSTSTKPGGEITTTFVTKNI 1065

Qy    1017 PRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGA 1076
           |       |||    :|:    ||: |   :  | |                |   |   |
Db    1066 PTTYLTTIAPTP--SVTTVTNFTPTT-ITTTVCSTGT------------NSAGETTSGC 1109

Qy    1077 EGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSDETNICNGKPVDGLTTLRNG 1136
             :|    :       | | :   |         ::    |  ||  ||   | ||
Db    1110 SPKTVTTTVPCSTGTGEYTTEATTLVTTAVTTTVVTTESSTGTNSA-GKTTTGYTTKSVP 1168

Qy    1137 TLVAFRGHYFWMLSPFSPPSPARRITEVWGIPSPIDTVFTRC----NCEGKT 1184
           |       |   |:| :| :||       :|: | |  | |    |  |:|
Db    1169 TT------YVTTLAPSAPVTPATN-----AVPTTITT--TECSAATNAAGET 1207


RESULT   7
S49915
extensin-like protein - maize
C;Species: Zea mays (maize)
C;Date: 05-Mar-1995 #sequence_revision 12-May-1995 #text_change 29-Oct-1999
C;Accession: S49915
R;Rubinstein, A.L.; Broadwater, A.H.; Lowrey, K.; Bedinger, P.A.
submitted to the EMBL Data Library, June 1994
A;Description: Pex genes: pollen-specific genes with extensin-like domains.
A;Reference number: S49915
A;Accession: S49915
A;Status: preliminary
A;Molecule type: DNA
A;Residues: 1-1188 <RUB>
A;Cross-references: EMBL:Z34465; NID:g600117; PIDN:CAA84230.1; PID:g600118


  Query Match             10.6%;  Score 784;  DB 2;  Length 1188;
  Best Local Similarity   29.9%;  Pred. No. 2.1e-25;
  Matches  230;  Conservative   59;  Mismatches  359;  Indels  122;  Gaps   33;

Qy     188 PTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPRPSL----PPN 243
           |||  |  |          |:   ||           | ||::        | |||    ||
Db     460 PTPHSPPAD---------DYVPPTPPVPGKSPPATSPSPQVQPPAASTPPPSLVKLSPPQ 510

Qy     244 SDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKV 303
           :                        |     ||||          |     | | |
Db     511 APVG------------------------SPPPPVKTTSPPAPIG----SPSPPPPVSVV 541

Qy     304 -----LAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPA 358
                :  | | |   :  |   :|  | |       |   :|  | || : | |   | |
Db     542 SPPPPVKSPPPPAPVGSPPPPEKSPPPPAPVASPPPPVKSP--PPPTLVASPPPPVKSPP 599

Qy     359 PTTTKSAPTTP--KEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTP-KEPAPTT 415
           |    ::|  |    | ||    | |  |   :|   | | | |  |:|  | | | |
Db     600 PPAPVASPPPPVKSPPPPTPVASPPPPAPVASSPPPMKSPPPPTPVSSPPPPEKSPPPPP 659

Qy     416 PKKPAPTTPKEPAPTT------PKE---PTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKP 466
           | |  |   : | | |       | |   | ||    | |   |     |:|| :| |::|:||
Db     660 PAKSTPPPEEYPTPPTSVKSSPPPEKSLPPPTLIPSPPPQEKPTPPSTPSKP-PSSPEKP 718

Qy     467 APTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPK 526
           :|  ||||  : |: |       | |     | ||   | |       :| : ||:|
Db     719 SP--PKEPVSSPPQTP----KSSPPPAPVSSPPPTPVSSPPALAPVSSPPSVKSSP---- 768

Qy     527 EPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAP 586
             |:|  ::  |||     | |     | |  ||   | |     |   | | ||    |
Db     769 PPAPLSSPPPAPQVKSSPPPVQVSSP-PPAPKSSPPLAPVSSPPQVEKTSPPPAPLSSPP 827

Qy     587 TTPKETAP----TTP----KKLTPTTPEKLAPTTPEKPAP-----TTPEELAPTTPEEPT 633
             || : |    ::|    |   |  |    | ||:  :|      : || : |:||  |
Db     828 LAPKSSPPHVVVSSPPPVVKSSPPPAPVSSPPLTPKPASPPAHVSSPPEVVKPSTP--PA 885

Qy     634 PTTPEEPAPTTPKAAAPNTPKE-PAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTT 692
           |||   | |: ||:: | ||   | |     | |     | | ||| : |     : | |
Db     886 PTTVISP-PSEPKSSPPPTPVSLPPPIVKSSPPPAMVSSP-PMTPKSSPPPVVVSSPPPT 943

Qy     693 LKE---PA-----PTTPKK---PAPKELAPTTTKE---PTSTTSDKPAPTTPKGTAPTTP 738
           :|    ||     | |||     |||  | |   |    ||  :|  ||   || : |  |
Db     944 VKSSPPPAPVSSPPATPKSSPPPAPVNLPPPEVKSSPPPTPVSSPPPA---PKSSPPPAP 1000
```

```
A;Accession: T30826
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-2187 <YOT>
A;Cross-references: EMBL:U48363; NID:g1666688; PID:g1666689; PIDN:AAB18732.1
C;Genetics:
A;Gene: Naca
A;Map position: 10
A;Introns: 24/1; 1996/1; 2024/3; 2050/3; 2099/3; 2142/3; 2183/3
A;Note: differential splicing converts alphaNAC into a tissue-specific DNA-binding activ
C;Keywords: alternative splicing; DNA binding; transcription factor

  Query Match             11.0%;  Score 812;  DB 2;  Length 2187;
  Best Local Similarity   27.4%;  Pred. No. 2.7e-26;
  Matches  311;  Conservative  128;  Mismatches  449;  Indels  246;  Gaps   51;

Qy      85 NPTSPPSSKKAPPPSGASQTIKSTTKRSPKPPNKKKTKKVIESEEITEEHSVSENQESSS 144
           :|| ||||| || ||  :          ||: |    |     |: :   :|   |
Db     746 SPT-PPSSKGAPVPSTGA-----------PPSPKGAPIVPTESSISSKQVPAEILPSPQ 792

Qy     145 SSSSSSSSSTIWKIKSSKNSAANRELQKKLKVKDNKKNRTKKKPTPKPPVVDEAGSGLDN 204
           :   ::|  |  ::| |      :     |      :|     ||
Db     793 KTPEVTASRLISAVQSPKVDPIMSD------VTPTSPKKTSATAVPK------------ 833

Qy     205 GDFKVTTPDTSTTQHNK-----VSTSPKITTAKPINPRPSLPPNSDTSKETSL--TVNKE 257
                 ||| |   |      | ||      | |    :|    || : :|     ||
Db     834 --------DTSATLSLKSVPAVTSLSPPKAPVAPSNEATIVPTEIPTSLKNALAAATPKE 885

Qy     258 T------TVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKA 311
           |      | :     | |  |  |    || | |   |  ::||::|:     |  |
Db     886 TLATSIPKVTSPSPQKTPKSVSLKGAPAMTSKKAT---EIAASKDVSPSQ--FPKEVPLL 940

Qy     312 ETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKE-PAPTTTKSAPTTPK 370
           :     |  : || |   |     :: | :  | |:   || ||: | |  :|  | ||
Db     941 QHV---PPTSPPKSPVSDTLSGALTSPPPKGPPATLAETPTYPKKSPKPAASKKTPATPS 997

Qy     371 ---------EPAPTTTKEPAPTTPKEPAPTTTKEPAPTT--TKSAPT------------- 406
                    |  | : | |    ||| : |:: : || |  :|  |:
Db     998 PEGVTAVPLEIPPCSKKAPKTAAPKESSATSSSKRAPKTAVSKEIPSKGVTAVPLEISLP 1057

Qy     407 --------TPKEPAPTTPKK-PAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKE 457
                   || | : ::||: |    |||   | |   |   |:   |  : |   || |
Db    1058 LKETSKSATPGEKSASSPKRSPKTAGPKE---TPPGGVTAVPPEISLPPKETPQNATPNE 1114

Qy     458 PAPTAPKKPAPTT--PKEPAP----TTPKEPAPTTTKEPSPTTPK-----EPAPTTTKSA 506
           : :| :| |  |||  |       | |      | |    ||     | | |    :
Db    1115 SLAASSQKRSPKTSVPKETPPGGVTAMPLEIPSAPQKAPKTAVPKQIPTPEDAVTILAGS 1174

Qy     507 PTTTKEPAPTTT-KSAPTTPK--------EPSPTTTKEPAPTTPKEPAPTTPKKPAPTT- 556
           | : |: : |    | || ||         | ||:  |       ||| : | | | :| |
Db    1175 PLSPKKASKTAAPKEAPATPSVGVIAVSGEISPSPKKTSKTAAPKENSATLPPKRSPKTA 1234

Qy     557 -PKE-----------PAPTTPKEPAP-------TTTKKPAPTAPKEPAPTTPKETAPTT 596
            |||            |:  :|  | |       | | | || |  | :| :||:   |
Db    1235 APKETPATSSEGVTAVPSEISPSPPTPASKGVPVTLTPKGAPNALAE-SPASPKKVPKTA 1293

Qy     597 PKKLTPTTPEKLAPTTPEK-PAPTTPEELAPTTPEEPTPTT--------PEEPAPTTPKA 647
             : | |||      :|:| |    |:| : | | : || |         | |      |
Db    1294 APEETSTTP------SPQKIPKVAGPKEASATPPSKKTPKTAVPKETSAPSEGVTAVPLE 1347

Qy     648 AAPNTPKEPAPTTPKE-PAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTP----- 701
             |:  | |    ||| |||:   | | | | :  |:  ||:     ||  ||||
Db    1348 IPPSPRKAPKTAAPKETPAPS--PEGATTAPVQIPPSPRKGSKKAGSKE-TPTTPSPEGV 1404

Qy     702 --------------KKPAPKELAPTTTKEPTS------TTSDKPAPTTPKGTAPTTPKE 740
                         |  :|||   | : :  |       || : |   |     |:  |
Db    1405 TAAPLEIPISSKKTSKMASPKETLVTPSSKKLSQTVGPKETSLEGATAVPLEIPPSHKKA 1464

Qy     741 PAPTTPKE-PAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTP 799
           |    ||: |   :||  ||||| | :|::||| |||   || : :  |:   :||| |
Db    1465 PKTVDPKQVPLTPSPK-DAPTTLAE-SPSSPKK-APKTAAPPSER-VTTVPPEKPA-TPQ 1519

Qy     800 KETAPTTPKEPAPTTPKKPAPTTPETP-----PPTTSEVSTPTTTK------EPTTIHKS 848
           | : |  | | |   :: | :: |||     ||  :  |   |:|       | |:  |
Db    1520 KASGTTASKVPVPAETQEVAVSSRETPVTPAVPPVKNPSSHKKTSKTIELKEAPATLPPS 1579

Qy     849 PDESTPELSAEPTPKALENSPKE-PGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETT 907
           | :|    |::  |:   ::||| |  |: | |  |   :  ||     :    | |:
Db    1580 PTKSPKIPSSKKAPRT--SAPKEFPASPSIK-PVTT--SLAQTAPPSLQKAPSTTIPKEN 1634

Qy     908 TAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKT 967
             ||| :       : |:  :   |: :   :|    | | : ||: :    |    | :|
Db    1635 LAAPAV----LPVSSKSPAAPARASASLSPATAAPQTAPKEATTIPSCKKAAATETPIET 1690

Qy     968 ITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPT 1027
             |   :   |:||::     : ||     |      ||| :|:|:  |:|
Db    1691 STAPSLEGAPKETSE----TSVSKVLMSSP------PKKASSSKRASTLP---------- 1730

Qy    1028 PRKMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETP 1081
                :|:| |   | ::         |  | : |   ::| | ||      ||
Db    1731 ----ATTLPSLKEASVLS---------PTATSSGKDSHISPVS-DACSTGTTTP 1770


RESULT   6
S48478
glucan 1,4-alpha-glucosidase (EC 3.2.1.3) - yeast (Saccharomyces cerevisiae)
N;Alternate names: extracellular glucoamylase; mucin-like protein MUC1; protein YIR01
C;Species: Saccharomyces cerevisiae
C;Date: 10-Sep-1999 #sequence_revision 10-Sep-1999 #text_change 12-Nov-1999
C;Accession: S48478; A26877; B26877; S27281; JC6123
R;Rowley, K.
submitted to the EMBL Data Library, October 1994
A;Reference number: S48478
A;Accession: S48478
A;Molecule type: DNA
A;Residues: 1-1367 <ROW>
A;Cross-references: GB:Z47047; EMBL:Z38061; NID:g603997; PID:g763364; GSPDB:GN00009;
R;Yamashita, I.; Nakamura, M.; Fukui, S.
J. Bacteriol. 169, 2142-2149, 1987
A;Title: Gene fusion is a possible mechanism underlying the evolution of STA1.
A;Reference number: A91831; MUID:87194600
A;Accession: A26877
A;Molecule type: DNA
A;Residues: 1-242 <YAM>
A;Cross-references: EMBL:M16164; NID:g172522; PIDN:AAA35014.1; PID:g172525
A;Accession: B26877
A;Molecule type: DNA
A;Residues: 762-1331 <YA2>
A;Cross-references: EMBL:M16165; NID:g172523; PIDN:AAA35015.1; PID:g172526
R;Pardo, J.M.; Ianez, E.; Zalacain, M.; Claros, M.G.; Jimenez, A.
FEBS Lett. 239, 179-184, 1988
A;Title: Similar short elements in the 5' regions of the STA2 and SGA genes from Sacc
A;Reference number: S27281; MUID:89031230
A;Accession: S27281
A;Molecule type: DNA
A;Residues: 1-31 <PAR>
A;Cross-references: EMBL:X13857; NID:g4551; PIDN:CAA32069.1; PID:g4552
R;Lambrechts, M.G.; Bauer, F.F.; Marmur, J.; Pretorius, I.S.
Proc. Natl. Acad. Sci. U.S.A. 93, 8419-8424, 1996
A;Title: Muc1, a mucin-like protein that is regulated by Mss10, is critical for pseud
A;Reference number: JC6123; MUID:96323237
A;Accession: JC6123
A;Status: nucleic acid sequence not shown
A;Molecule type: DNA
A;Residues: 1-1367 <LAM>
A;Cross-references: GB:U30626; NID:g1304386; PIDN:AAC49609.1; PID:g1304387
C;Genetics:
A;Gene: SGD:MUC1; STA2; MAL5; DEX2; SGD:S0001458
A;Cross-references: MIPS:YIR019c; SGD:S0001458
A;Map position: 9R
C;Superfamily: yeast glucan 1,4-alpha-glucosidase homolog; glucan 1,4-alpha-glucosida
C;Keywords: glycosidase; hydrolase; polysaccharide degradation; transmembrane protein
F;5-21/Domain: transmembrane #status predicted <TM1>
F;1350-1366/Domain: transmembrane #status predicted <TM2>
```

```
Db    988 ETTYAPTEETMYAPIEETTYAPTEETTYAPAEETPYEPTEETTYAPTEETTYAPTEETTY 1047

Qy    761 ------------TLKEPAPTTPKKP------APKE---LAPT--TTKGPTSTTSDKPAP 796
Db   1048 ASTEETTYAPTEETTYAPAEETPYEPTEETTYAPTEETTYAPTEETTYAPTEETTYAPTE 1107

Qy    797 TTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPT--TIHKSPDESTP 854
Db   1108 ETTYAPAEETPYEPTEETTYAPTEETTYAPTEETMYAPIEETTYGPTEETTY-APTEATT 1166

Qy    855 ELSAEPTPKA-LENSPKEPG-----VPTTKTPAATKPEMTTTAKDKTTERDLRTTP---- 904
Db   1167 YAPTEETPYAPTEETTYEPTGETTYAPTEETTYAPTEETTYAPTEETTYAPTEETPYEPT 1226

Qy    905 ETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTT----------PFKITT--- 951
Db   1227 EETTYAP-----TEETTYEPTEETTYAPTEETTYAPTEETTYAPTEETMYAPIDETTYGP 1281

Qy    952 LKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRAT--NSKATTPKP-QKPTKAPKKPT 1008
Db   1282 TEETTYAPTEATTYAPTEETPYAPTEETTYEPTGETTYAPTEETTYAPTEETTYAPMEET 1341

Qy   1009 ------------STKKP-KTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRP 1055
Db   1342 PYEPAEESTSTVSTEKPCNTEEFTDEPTDEPTDE--PSDEPTDEPTDEPTDLPTDEPSTP 1399

Qy   1056 --NQTPNSKLVEVNPKSEDAG 1074
Db   1400 CDNQGINGIGVENKVRYNNAG 1420
```

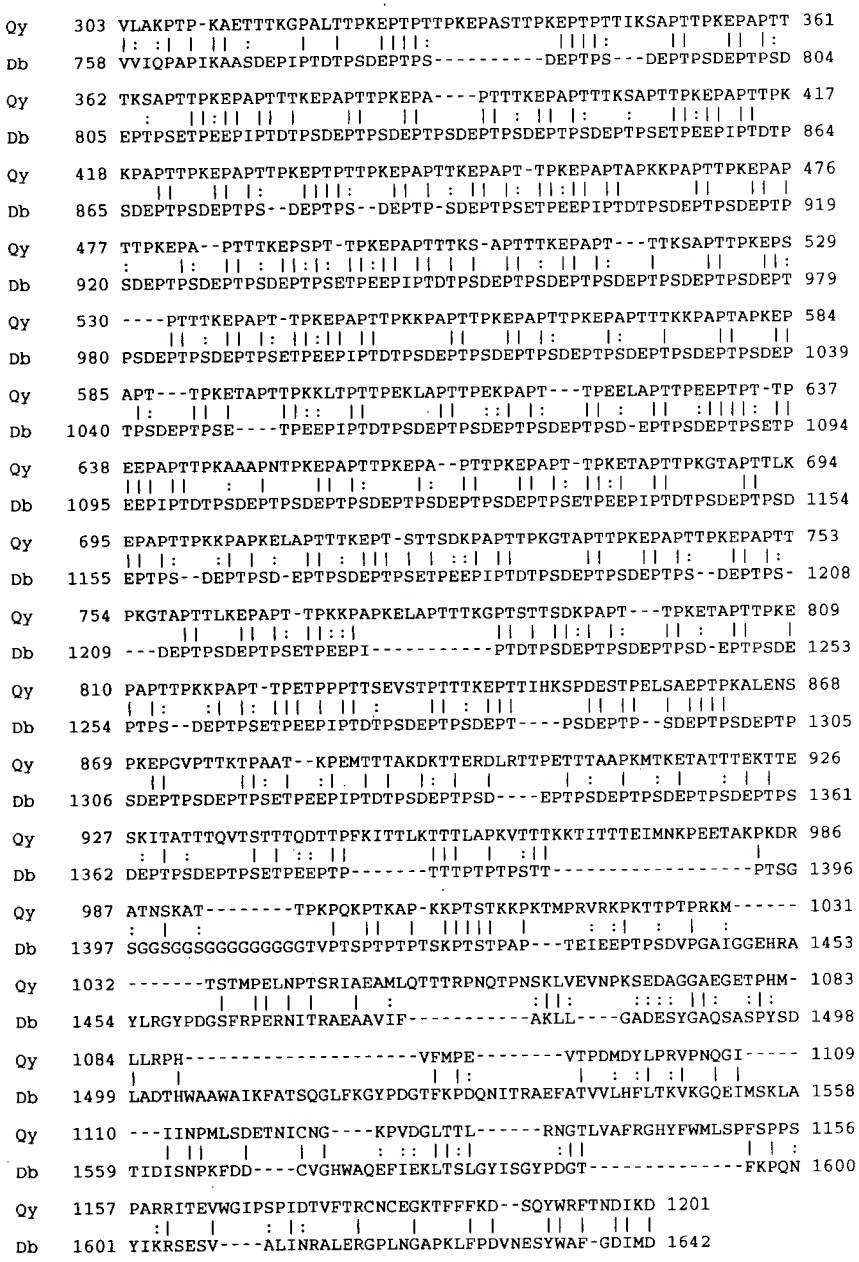
RESULT 4
T16251
hypothetical protein F35A5.1 - Caenorhabditis elegans
C;Species: Caenorhabditis elegans
C;Date: 20-Sep-1999 #sequence_revision 20-Sep-1999 #text_change 18-Feb-2000
C;Accession: T16251
R;Leimbach, D.
submitted to the EMBL Data Library, January 1996
A;Description: The sequence of C. elegans cosmid F35A5.
A;Reference number: Z18485
A;Accession: T16251
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-1274 <LEI>
A;Cross-references: EMBL:U46675; NID:g1166613; PID:g1166621; PIDN:AAB52641.1; GSPDB:GN00
A;Experimental source: strain Bristol N2; clone F35A5
C;Genetics:
A;Gene: CESP:F35A5.1
A;Map position: X
A;Introns: 1272/2

Query Match 11.4%; Score 846; DB 2; Length 1274;
Best Local Similarity 27.6%; Pred. No. 6.5e-28;
Matches 313; Conservative 119; Mismatches 455; Indels 248; Gaps 60;

```
Qy     90 PSSKKAPPPSGASQTIKSTTKRSP----KPP---NKKKTKKVIESEEITEEHSVSENQES 142
Db    185 PSPKKAAPSKEHDPIVPPTPIKNPAKKWKPPWEDDEVPTEEIKEPEPATRKVPALKKKEP 244

Qy    143 SSSSSSSSSSSTIWKIKSSKNSAANRELQKKLKVKDNK-KNRTKK-KP-----TP----- 190
Db    245 STSVKPVSDPSPTKKV----------PVKKEPEVPPTPIKNPTKKWKPPWEDETPVEEVK 294

Qy    191 KPPVVDEAGSGLDNGD----FKVTTPDTSTTQHNKVSTSPKITTAKPI-NPRPSLPPNSD 245
Db    295 EPPVPEKKAPVLKKKDPAPAAKARDPSPSKAAPKKVEPSSPVVPPTPVKNPVKKYKPPWE 354

Qy    246 TSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLA 305
Db    355 VDDEPAEEVKKPSAPEKKTPVLKRKEPEPSSTTPSSDPSPKKAAPAVKPRDSSPKKATPL 414

Qy    306 KPTPKAE----TTTKGPA----------LTTPKEPTPTTP----KEPASTTPK 340
Db    415 QADPKAQEVPPTPVKNPVKKYKPPWEVDDEDPVEEVKQPEAPAKKTPVLKRKEPAAKDTA 474

Qy    341 EP----TP-TTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPA---------- 381
Db    475 KPATSKTPETPEKKDPVKPRDSSPKKVAAKPDSAQAPA-TPVKNPVKKWRPPWEDDETPA 533

Qy    382 -----PTTPKEPAPTTTKEPAPT-------TTKSAPTTPKEP-----APTTPKKPAPTTP 424
Db    534 DDVSKPTDAKKTPSLAKKDPAPAKESLKPKADTKAPAKPRDPSPKKVAPTAPEKKTPVLA 593

Qy    425 KE----PAPTTPKEPTPTTPKEPAPTTKEPAPTTPK-EPAPTAPKKPAP------TTPKE 473
Db    594 KKEPAGPADSKTKEPEKSKPRDPSPKKAVPAKPVPKTEVAPAAVKKPEPISKPKDTAPKK 653

Qy    474 PAPTTPKEPAPTTTKEP----SPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTP-KEP 528
Db    654 AEPNSPVVP-PTPVKNPVKKWKPPWEDDDAPAKPVSLPEPEKK-TPVLAKKAPTKPDSEA 711

Qy    529 SPTTTKEPAPTTPK--EPAPTTPKKPAP--TTPKEPAPTTPKEPAPTTTKKPAPTA-PKE 583
Db    712 AADPVSGPSSKDPKLAKKAPVKPRDPSPMKAVPIKPAPKT--EVPPAVVKKPEPVAKSRD 769

Qy    584 PAPTTPKETAPTTPKKLTPTTP-----EKLAPTTPEKPAPTTPEELAPTTPEEP-----T 633
Db    770 PSPKKAK-AEPNSP--VVPPTPVKNPVKKWKPPWEDDDAPAEPVNVPEPEKKTPVLAKKT 826

Qy    634 PTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTT-PKEPAPTTPKETAPTTPKGTAPTT 692
Db    827 PVKPRDPSPKKAVPAKPSTKTDAPPVSVKKPEPVSKPKEPSPKKAEPNSPVVP----PTP 882

Qy    693 LKEPAPTTPKKPAPKELAPT-TTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAP 751
Db    883 VKNPVKKW-KPPWEDDDEPTEEVKKPSE--PEKKTPVLAK-KEPEKPKD-APKVAAKPRD 937

Qy    752 TTPKGTAPTTLKEPAPT-------TPKKPAP-----KELAPTTTKGPTS------TTSDK 793
Db    938 PSPKKAVPE--KEPAKVAAKPRDLSPKKAIPIPANTQEAPPTPVKNPVKKWKPPWEDDDE 995

Qy    794 P-----APTTPKET------APTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEP 842
Db    996 PAEPVSAPEPEKKTPVLAKKAPAKPRDP---SPKKAAPVAAK-PDPKIPEV-PPTPVKNP 1050

Qy    843 TTIHKSP----DESTPELSA-EP---TPKALENSPKEPGV------------PTTKTPA 881
Db   1051 VKKWKPPWEDDDEPSEPVSAPEPEKKTPVLAKKAPTKPATKPDSEAAADPVSGPTSKDPK 1110

Qy    882 AT------KPEMTTTAKDKTTERDLRTTPETT--TAAPKMTKETATTTEKTTESKITATT 933
Db   1111 LSKKAPVEKPKPTTDPKDDKLKPSPAKKPEKAPEPAAPKKWKPVWDDDPDEPEADFTVPA 1170

Qy    934 TQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKAT 993
Db   1171 PSKKPDTEDPADPLG---------GPKTKDPK--------LNKKAPAEKPTEK------- 1206

Qy    994 TPKPQKPTKAPKKPTSTKKPKTMPRVRKPK-----------TTPTPRKMTSTMP 1036
Db   1207 -PKPKEVSKEPPKPTEPPKP-AAPKKWKPPWEDDPDEPEADFTMPAPKKPDTEDP 1259
```

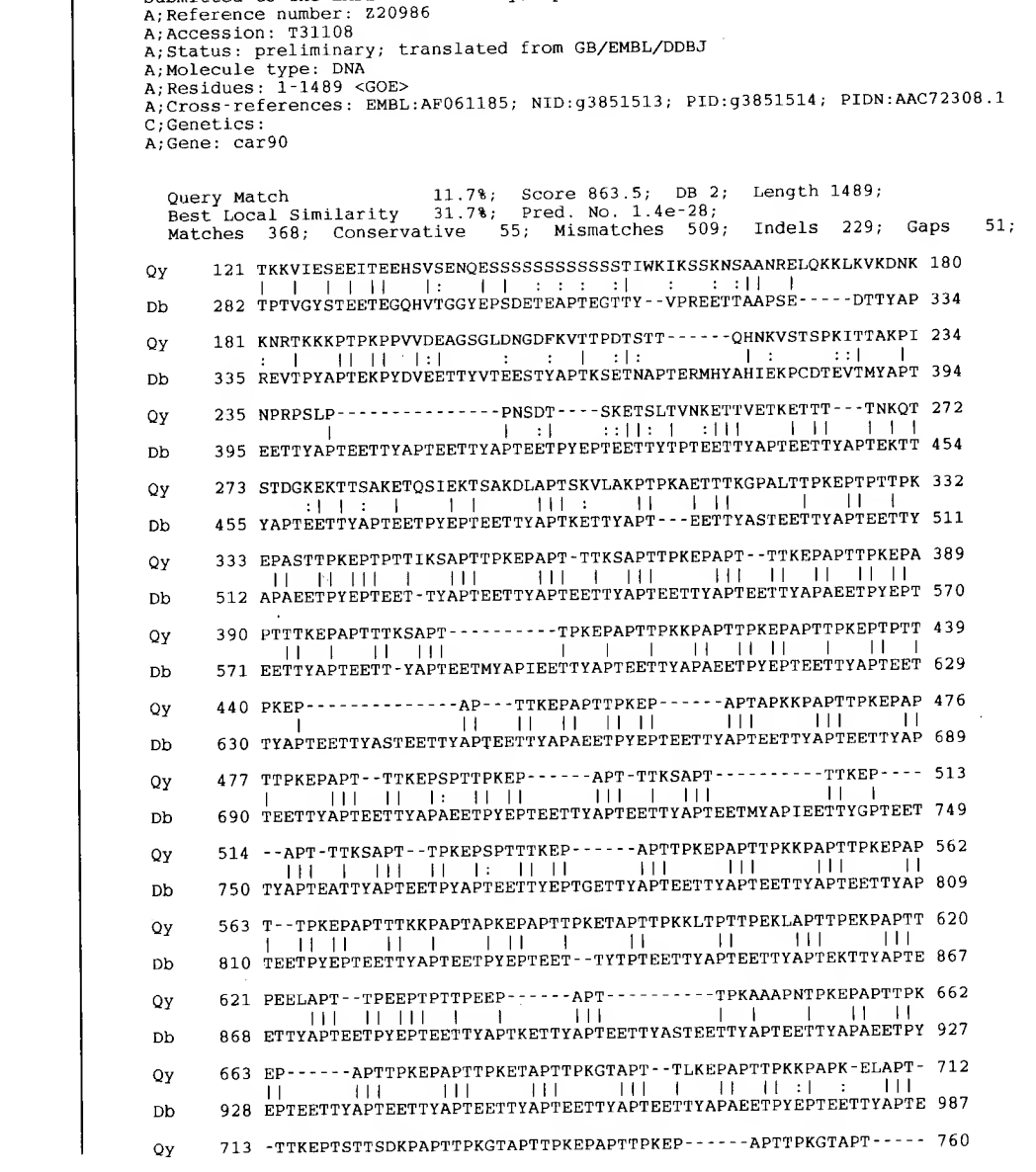
RESULT 5
T30826
nascent polypeptide-associated complex alpha chain, muscle splice form gp220 - mouse
N;Alternate names: alpha-NAC protein
C;Species: Mus musculus (house mouse)
C;Date: 22-Oct-1999 #sequence_revision 22-Oct-1999 #text_change 24-Nov-1999
C;Accession: T30826
R;Yotov, W.V.; St-Arnaud, R.
Genes Dev. 10, 1763-1772, 1996
A;Title: Differential splicing-in of a proline-rich exon converts alphaNAC into a mus
A;Reference number: Z20889; MUID:96312450

Query Match 12.8%; Score 950; DB 2; Length 1664;
Best Local Similarity 31.0%; Pred. No. 4.5e-32;
Matches 312; Conservative 104; Mismatches 361; Indels 230; Gaps 55;



RESULT 3
T31108
cyst germination specific acidic repeat protein precursor - Phytophthora infestans
C;Species: Phytophthora infestans (potato late blight agent)
C;Date: 22-Oct-1999 #sequence_revision 22-Oct-1999 #text_change 22-Oct-1999
C;Accession: T31108
R;Goernhardt, B.
submitted to the EMBL Data Library, April 1998
A;Reference number: Z20986
A;Accession: T31108
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-1489 <GOE>
A;Cross-references: EMBL:AF061185; NID:g3851513; PID:g3851514; PIDN:AAC72308.1
C;Genetics:
A;Gene: car90

Query Match 11.7%; Score 863.5; DB 2; Length 1489;
Best Local Similarity 31.7%; Pred. No. 1.4e-28;
Matches 368; Conservative 55; Mismatches 509; Indels 229; Gaps 51;



J. Clin. Invest. 87, 77-82, 1991
A;Title: Human bronchus and intestine express the same mucin gene.
A;Reference number: A61257; MUID:91086481
A;Accession: A61257
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 'T',1925-1948,'TTS',1952-1954 <JAN>
A;Experimental source: bronchus
R;Xu, G.; Huan, L.; Khatri, I.; Sajjan, U.S.; McCool, D.; Wang, D.; Jones, C.; Forstner,
Biochem. Biophys. Res. Commun. 183, 821-828, 1992
A;Title: Human intestinal mucin-like protein (MLP) is homologous with rat MLP in the C-t
A;Reference number: PQ0328; MUID:92198477
A;Accession: PQ0328
A;Molecule type: mRNA
A;Residues: 2328-2468 <XUG>
A;Cross-references: GB:M86523
A;Experimental source: small intestine
A;Accession: PQ0329
A;Molecule type: protein
A;Residues: 2328-2342,'K',2344-2354 <XUG1>
C;Genetics:
A;Gene: GDB:MUC2
A;Cross-references: GDB:120203; OMIM:158370
A;Map position: 11p15.5-11p15.5
C;Superfamily: von Willebrand factor; von Willebrand factor type A repeat homology; von
C;Keywords: glycoprotein; intestine; tandem repeat
F;2766-2834/Domain: von Willebrand factor type C repeat homology <VWC>

Query Match 14.7%; Score 1090; DB 2; Length 3020;
Best Local Similarity 27.4%; Pred. No. 1.4e-37;
Matches 378; Conservative 107; Mismatches 514; Indels 382; Gaps 51;

```
Qy      19 TCNCDY--NCQHYMECCPDFKRVCTAELSCKGRCFESFERGRECDCDAQCKKYDKCCPDY 76
Db    1079 SCSCDTGGDCECFCSAVASYAQECTKEGAC------VFWRTPDL-CPIFCDYYN---PPH 1128

Qy      77 ESFCAEVHNPTSPPSSKKAPPPSGASQTI-----KSTTKRSPK--PPNKKKTKKVIESEE 129
Db    1129 E--CEWHYEPCGNRSFETCRTINGIHSNISVSYLEGCYPRCPKDRPIYEEDLKKCVTADK 1186

Qy     130 ----ITEEH----SVSENQESSSSSSSSSSSTIWKIKSSKNSAANRELQKKLKVKDNKK 181
Db    1187 CGCYVEDTHYPPGASVPTEETCKSCVCTNSSQVVCRPEEGK----------------IL 1229

Qy     182 NRTKKKPTPKPPVVDEAGSGLDNGDF---KVTTPDTSTTQH---NKVSTSPKITTAKPIN 235
Db    1230 NQTQ-----------------DGAFCYWEICGPNGTVEKHFNICSITTRPSTLTTFTTI 1271

Qy     236 PRPSLPPNSDTSKETSLTVNKETTVETKET--TTTNK--------------QTSTDGKE 278
Db    1272 TLPTTP--------TSFTTTTTTTTPTSSTVLSTTPKLCCLWSDWINEDHPSSGSDDGDR 1323

Qy     279 KTTS--AKETQSIEKTSAKD---------------------------------------- 296
Db    1324 EPFDGVCGAPEDIECRSVKDPHLSLEQHGQKVQCDVSVGFICKNEDQFGNGPFGLCYDYK 1383

Qy     297 ------------LAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTP 344
Db    1384 IRVNCCWPMDKCITTPSPPTTTPSPPPTTTTTLPPTTTPSPPTTTTTTPPPTTTPSPPIT 1443

Qy     345 TTIKSAP-TTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKS 403
Db    1444 TTTTPLPTTTPSPPISTTTTPPPTTTPSP-PTTTPSPPTTTPSPPT-TTTTTPPPTTTPS 1501

Qy     404 AP-TTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTA 462
Db    1502 PPMTTPITPPASTTTLPPTTTPSPPTTTTTTPPPTTTPSPPTTTPITP-PTSTTTLPPTT 1560

Qy     463 PKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAP-TTTKEPAPTTTKSA 521
Db    1561 TPSPPPTTTTTPPPTTTPSPPTTTTPSPPTITTTTPPPTTTPSPPTTTTTTPPPTTTPSP 1620

Qy     522 P-TTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTA 580
Db    1621 PTTTPITPPTSTTTLPPTTTPSPPPTTTTTPPPTTTPSPPTTTTPSPPITTTTTPPPTTT 1680

Qy     581 PKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEP 640
Db    1681 PSSPITTTP--SPPTT----TMTTPS--PTTTPSSPITTTTTPSSTTTP-SPPPTTMTTP 1731

Qy     641 APTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTP-----KETAPTTP---------- 685
Db    1732 SPTTTPSPPTTTMTTLPPTTTSSPLTTTPLPPSITPPTFSPFSTTTPTTPCVPLCNWTGW 1791

Qy     686 --KG------------------------------------------------------- 687
Db    1792 LDSGKPNFHKPGGDTELIGDVCGPGWAANISCRATMYPDVPIGQLGQTVVCDVSVGLICK 1851

Qy     688 ---------------------------TAPTTLKEPAPTTPKKPAPKELAPTT--TKEP 717
Db    1852 NEDQKPGGVIPMAFCLNYEINVQCCECVTQPTTM---TTTTTENPTPTPITTTTTVTPTP 1908

Qy     718 TSTTSDKP----APT-TPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKK 772
Db    1909 TPTSTQSPNGLQAPTPTPISTTTTVTPTPTPTGTQTPT-TTPITTTTTVTPTPTPTGTQT 1967

Qy     773 PAPKELAPTTTKGPTST-TSDKPAPTTPKETAPTTPKEPAPTTPKKP-------APTTPE 824
Db    1968 PTTVLITTTTTMTPTPTPTSTKSTTVTPITTTTTVTATPTPTGTQTPTMIPISTTTTVTP 2027

Qy     825 TPPPTTSEVSTPTTTK-----EPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKT 879
Db    2028 TPTPTTGSTGPPTHTSTAPIAELTTSNPPPESSTPQTSRSTSSPLTESTTLLSTLPPAIE 2087

Qy     880 PAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTST 939
Db    2088 MTSTAPPSTPTAPTTTSGGHTLSPPPSTTTSPPGTPTRGTTT--GSSSAPTPSTVQTTTT 2145

Qy     940 TTQDTTPFKITT---LKTTTLAPKVTTTKKTITTTEIMNKPEE---------------- 979
Db    2146 SAWTPTPTPLSTPSIIRTTGLRPYPSSVLICCVLNDTYYAPGEEVYNGTYGDTCYFVNCS 2205

Qy     980 ------------TAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPT 1027
Db    2206 LSCTLEFYNWSCPSTPSPTPTPSK-STPTPSKPSSTPSKPTPGTKPPECPDFDPPR---- 2260

Qy    1028 PRKMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNP------------KSEDAG 1074
Db    2261 ---------QENETWWLCDCFM-ATCKYNNTVEIVKVECEPPPMPTCSNGLQPVRVEDPD 2310

Qy    1075 G------------AEGETPHMLL----------RPHVFMPEVTPDMDYLPRVPNQGIII 1111
Db    2311 GCCWHWECDCYCTGWGD-PHYVTFDGLYYSYQGNCTYVLVEEISPSVD------NFGVYI 2363

Qy    1112 N 1112
Db    2364 D 2364
```

RESULT 2
T18262
S-layer protein - Clostridium thermocellum
C;Species: Clostridium thermocellum
C;Date: 15-Oct-1999 #sequence_revision 15-Oct-1999 #text_change 15-Oct-1999
C;Accession: T18262
R;Fujino, T.; Beguin, P.; Aubert, J.P.
J. Bacteriol. 175, 1891-1899, 1993
A;Title: Organization of a Clostridium thermocellum gene cluster encoding the cellulo
e.
A;Reference number: Z18847; MUID:93209931
A;Accession: T18262
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-1664 <FUJ>
A;Cross-references: EMBL:X67506; NID:g296879; PID:g296881; PIDN:CAA47841.1

GenCore version 4.5
Copyright (c) 1993 - 2000 Compugen Ltd.

OM protein - protein search, using sw model

Run on: April 26, 2002, 16:18:10 ; Search time 82.85 Seconds
(without alignments)
1268.810 Million cell updates/sec

Title: US-09-556-246-1_COPY_25_1404
Perfect score: 7410
Sequence: 1 QDLSSCAGRCGEGYSRDATC.....ARAITTRSGQTLSKVWYNCP 1380

Scoring table: BLOSUM62
Gapop 10.0 , Gapext 0.5

Searched: 219241 seqs, 76174552 residues

Total number of hits satisfying chosen parameters: 219241

Minimum DB seq length: 0
Maximum DB seq length: 2000000000

Post-processing: Minimum Match 0%
Maximum Match 100%
Listing first 45 summaries

Database : PIR_68:*
1: pir1:*
2: pir2:*
3: pir3:*
4: pir4:*

Pred. No. is the number of results predicted by chance to have a score greater than or equal to the score of the result being printed, and is derived by analysis of the total score distribution.

SUMMARIES

Result No.	Score	% Query Match	Length	DB	ID	Description
1	1090	14.7	3020	2	A43932	mucin 2 precursor,
2	950	12.8	1664	2	T18262	S-layer protein -
3	863.5	11.7	1489	2	T31108	cyst germination s
4	846	11.4	1274	2	T16251	hypothetical prote
5	812	11.0	2187	2	T30826	nascent polypeptid
6	809	10.9	1367	1	S48478	glucan 1,4-alpha-g
7	784	10.6	1188	2	S49915	extensin-like prot
8	695	9.4	1151	2	T18535	high molecular mas
9	676.5	9.1	1229	2	T25697	hypothetical prote
10	672	9.1	1344	1	A35175	mucin 1 precursor,
11	666.5	9.0	7962	2	I38346	elastic titin - hu
12	663	8.9	3570	2	T45025	mucin MUC5B, trach
13	658.5	8.9	3507	2	T34513	hypothetical prote
14	643.5	8.7	990	2	I51618	nucleolar phosphop
15	633	8.5	489	2	T11622	extensin class 1 p
16	632	8.5	761	2	C84672	hypothetical prote
17	629	8.5	6642	2	T29757	protein UNC-89 - C
18	627.5	8.5	971	2	T19431	hypothetical prote
19	622.5	8.4	839	2	F75518	hypothetical prote
20	607.5	8.2	801	2	T29018	hypothetical prote
21	607.5	8.2	924	2	S27923	gene LF3 protein -
22	605	8.2	379	2	S50125	larval glue protei
23	591.5	8.0	2232	2	T34434	hypothetical prote
24	568.5	7.7	1630	2	A53577	ascites sialoglyco
25	559.5	7.6	350	2	S22456	hydroxyproline-ric
26	559	7.5	875	2	S23760	polyphenolic adhes
27	556.5	7.5	856	2	T16543	hypothetical prote
28	551	7.4	620	2	S06733	hydroxyproline-ric
29	551	7.4	1087	1	QFMSH	neurofilament trip
30	545	7.4	873	2	A47283	calphotin - fruit
31	542	7.3	369	2	S20500	hydroxyproline-ric
32	540	7.3	1459	2	T32271	hypothetical prote
33	538.5	7.3	416	2	JU0465	extensin precursor
34	537	7.2	1072	1	A37221	neurofilament trip
35	532	7.2	756	2	T27642	hypothetical prote
36	530.5	7.2	813	2	S70795	vsaA protein precu
37	530.5	7.2	1162	2	JH0557	exo-alpha-sialidas
38	530	7.2	865	2	A47282	calcium-binding pr
39	522	7.0	328	2	JQ0985	hydroxyproline-ric
40	521.5	7.0	854	2	S02003	neurofilament trip
41	518	7.0	866	2	T45462	membrane glycoprot
42	518	7.0	1611	2	T38236	hypothetical prote
43	515.5	7.0	3534	2	T42567	tegument protein 2
44	513	6.9	1832	2	T31113	mucin-like glycopr
45	512.5	6.9	867	2	T45463	membrane glycoprot

ALIGNMENTS

RESULT 1
A43932
mucin 2 precursor, intestinal - human (fragments)
N;Alternate names: mucin SMUC-41
C;Species: Homo sapiens (man)
C;Date: 10-Mar-1993 #sequence_revision 12-Apr-1996 #text_change 05-Nov-1999
C;Accession: A49963; A45106; B45106; A43932; B33532; A61257; PQ0328; PQ0329
R;Gum Jr., J.R.; Hicks, J.W.; Toribara, N.W.; Siddiki, B.; Kim, Y.S.
J. Biol. Chem. 269, 2440-2446, 1994
A;Title: Molecular cloning of human intestinal mucin (MUC2) cDNA. Identification of t
A;Reference number: A49963; MUID:94132002
A;Accession: A49963
A;Molecule type: mRNA
A;Residues: 1-639 <GU1>
A;Cross-references: GB:L21998
R;Gum Jr., J.R.; Hicks, J.W.; Toribara, N.W.; Rothe, E.M.; Lagace, R.E.; Kim, Y.S.
J. Biol. Chem. 267, 21375-21383, 1992
A;Title: The human MUC2 intestinal mucin has cysteine-rich subdomains located both up
A;Reference number: A45106; MUID:93016075
A;Accession: A45106
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 626-1895 <GU2>
A;Cross-references: GB:M94131; NID:g186395; PIDN:AAA59163.1; PID:g186396
A;Note: sequence extracted from NCBI backbone (NCBIP:116706)
A;Accession: B45106
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 2037-3020 <GU3>
A;Cross-references: GB:M94132; NID:g186397; PIDN:AAA59164.1; PID:g186398
A;Experimental source: colon
A;Note: sequence extracted from NCBI backbone (NCBIP:116698)
R;Toribara, N.W.; Gum Jr., J.R.; Culhane, P.J.; Lagace, R.E.; Hicks, J.W.; Petersen,
J. Clin. Invest. 88, 1005-1013, 1991
A;Title: MUC-2 human small intestinal mucin gene structure. Repeated arrays and polym
A;Reference number: A43932; MUID:91358717
A;Accession: A43932
A;Molecule type: DNA
A;Residues: 1343-1350,'L',1352-1411,'S',1413-1448,'P',1450-1503,'T',1505-1915 <TOR>
A;Cross-references: GB:M74027; NID:g188863; PIDN:AAA59875.1; PID:g188864
A;Note: sequence inconsistent with the nucleotide translation
A;Note: sequence extracted from NCBI backbone (NCBIN:55749, NCBIP:55750)
R;Gum, J.R.; Byrd, J.C.; Hicks, J.W.; Toribara, N.W.; Lamport, D.T.A.; Kim, Y.S.
J. Biol. Chem. 264, 6480-6487, 1989
A;Title: Molecular cloning of human intestinal mucin cDNAs. Sequence analysis and evi
A;Reference number: A33532; MUID:89197956
A;Accession: B33532
A;Molecule type: mRNA
A;Residues: 1916-2193 <GU4>
A;Cross-references: GB:M22405; NID:g188873; PIDN:AAA36334.1; PID:g188874
A;Experimental source: intestine
R;Jany, B.H.; Gallup, M.W.; Yan, P.S.; Gum, J.R.; Kim, Y.S.; Basbaum, C.B.

```
Db    821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy    824 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPT-SRIAEAMLQTTTRPNQ 882

Db    869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPXPGSTAPXAHGVTSAPDX 924

Qy    883 TP 884

Db    925 RP 926
```

RESULT 15
US-08-276-967-2
; Sequence 2, Application US/08276967
; Patent No. 5851817
; GENERAL INFORMATION:
; APPLICANT: Hardy, Daniel M.
; APPLICANT: Garbers, David L.
; TITLE OF INVENTION: Species-Specific Egg-Binding Proteins of
; TITLE OF INVENTION: Sperm
; NUMBER OF SEQUENCES: 8
; CORRESPONDENCE ADDRESS:
; ADDRESSEE: Arnold, White & Durkee
; STREET: P. O. Box 4433
; CITY: Houston
; STATE: Texas
; COUNTRY: USA
; ZIP: 77210-4433
; COMPUTER READABLE FORM:
; MEDIUM TYPE: Floppy disk
; COMPUTER: IBM PC compatible
; OPERATING SYSTEM: PC-DOS/MS-DOS
; SOFTWARE: PatentIn Release #1.0, Version #1.25
; CURRENT APPLICATION DATA:
; APPLICATION NUMBER: US/08/276,967
; FILING DATE: Submitted Herewith
; CLASSIFICATION: 530
; ATTORNEY/AGENT INFORMATION:
; NAME: Kitchell, Barbara S.
; REGISTRATION NUMBER: 33,928
; REFERENCE/DOCKET NUMBER: UTSD:418\KIT
; TELECOMMUNICATION INFORMATION:
; TELEPHONE: 713-787-1400
; TELEFAX: 713-789-2679
; TELEX: 79-0924
; INFORMATION FOR SEQ ID NO: 2:
; SEQUENCE CHARACTERISTICS:
; LENGTH: 2476 amino acids
; TYPE: amino acid
; STRANDEDNESS: single
; TOPOLOGY: linear
; MOLECULE TYPE: protein
US-08-276-967-2

Query Match 8.4%; Score 419.5; DB 2; Length 2476;
Best Local Similarity 32.5%; Pred. No. 8.5e-19;
Matches 163; Conservative 43; Mismatches 177; Indels 119; Gaps 31;

```
Qy    406 PKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEP 465

Db    313 PSETSVSTEKPVAPTE----KPTVPSEIYTIPTEKPMVHMEKPIVHT--EKPTVPT-EKP 365

Qy    466 APTTPKAAAPNTPKEPA----PTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEP 521

Db    366 TIPTEKSTVPT--KKPTVFKEPTLPPE-GPTVPAE-RPTTPPEGPAVPPKG--PTVLTE- 418

Qy    522 APTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPT-TPKEPAPTTPKEPAPTTPK 580

Db    419 ------------WPTSHTEKSTVHTEKPILPTGKSTIPTEKPMVPTKRT------TTP- 458

Qy    581 GTAPTTLKEPAPTTPKKPAPKELAPT-TTKGPTSTT---SDKPAPTTPKETAPTTPKEPA 636

Db    459 -TERTTIPAEKPTV---PIEKPMVPTERTTIPTERTTIPTEKPTVPTEKLTVPT--EKPI 512

Qy    637 PTTPKKPAPTTPETPPP-----TTSEVSTPT-----TTKEPTTIHKSPDESTPELSAEPT 686

Db    513 VPTEKPIVPTEKHTIPTEKLTVLTERTTTPTERTTIPTEKPTVPTEKPSVPT-EKPTVPT 571

Qy    687 PKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTT 746

Db    572 --------EEPTIPTEKLTVPT--ERTTTPTKRTTTPTIRTTTPTI--------RTTTPT 613

Qy    747 EKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETA 806

Db    614 ERTT------TPTIRTTTPTERTT---IPTKKTT------VPTEKTIIPT------ERTI 652

Qy    807 KPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPT 866

Db    653 AP---------TTPQP-SPTLVPTQPAAVVMPST------SATTVTPRTTIASCP---PN 693

Qy    867 SRIAEAMLQTTTRPNQTPNSKL 888

Db    694 AHFERCACPVSCQ-SPTPNCEL 714
```

Search completed: April 26, 2002, 16:16:35
Job time: 186 sec

```
; Sequence 2, Application US/09083116
; Patent No. 6203795
;  GENERAL INFORMATION:
;    APPLICANT:  CHAMBON, Pierre
;    APPLICANT:  KIENY, Marie-Paule
;    APPLICANT:  LATHE, Richard
;    APPLICANT:  HAREUVENI, Mara
;    TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;    TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
;    NUMBER OF SEQUENCES:  5
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
;      STREET:  P.O. Box 1404
;      CITY:  Alexandria
;      STATE:  Virginia
;      COUNTRY:  United States
;      ZIP:  22313-1404
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/09/083,116
;      FILING DATE:
;      CLASSIFICATION:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  08/479,537
;      FILING DATE:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  WO PCT/FR91/00835
;      FILING DATE:  23-OCT-1991
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/039,320
;      FILING DATE:  04-APR-1993
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/403,576
;      FILING DATE:  14-MAR-1995
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Teskin, Robin L.
;      REGISTRATION NUMBER:  35,030
;      REFERENCE/DOCKET NUMBER:  017753-025
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (703) 836-6620
;      TELEFAX:  (703) 836-2021
;  INFORMATION FOR SEQ ID NO:  2:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  2035 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  peptide
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  128..1899
;      OTHER INFORMATION:  /note= "The amino acids spanning
;      OTHER INFORMATION:  128 to 1899 constitute a repeated region wherein the repeat i
;      OTHER INFORMATION:  20 amino acids, 17 of which are fixed. The number of such
;      OTHER INFORMATION:  repeats varies from 1 to 40."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  134
;      OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;      OTHER INFORMATION:  Xaa Xaa which is the codon for Pro or Ala wherein Pro = CCT,
;      OTHER INFORMATION:  CCC, CCA, or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  144
;      OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;      OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC, ACA
;      OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  147
;      OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC,
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  1..21
;      OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;      OTHER INFORMATION:  21 amino acid precursor sequence."
US-09-083-116-2


  Query Match             8.5%;  Score 423.5;  DB 4;  Length 2035;
  Best Local Similarity   24.4%;  Pred. No. 3.9e-19;
  Matches  235;  Conservative   63;  Mismatches  461;  Indels  203;  Gaps   48;

Qy      75 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 133
           | |||           | :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db      16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy     134 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 176
           |    :||:|    |    ||               | |     ||||     |   |||
Db      66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy     177 TTPKEPAPTTT-------KSAPTTPKEP-----------APTTTKEPAPTTPKEPAPTT 217
           :||| :|          |||      |           ||      | | |      |:
Db     123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy     218 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 264
           :  | | :|          |||      | | |        ||      | | |      |:
Db     181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy     265 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 311
           :  | | :  |         ||      | |||        ||      | | |      |:
Db     241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy     312 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 360
           :  |  |   ||       |||      |  |          |||      |  |
Db     301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy     361 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 412
           ||      | | |      ||      | | |        ||      |  |||      |
Db     360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy     413 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 468
           :  :  |    | |      |:     ||      | | |      |:  :  |      | |
Db     420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy     469 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 523
           | |     :||:    |  | |      |:  :  |  |  |||      | ||      |
Db     472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy     524 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 571
           | |           :| |   ||       ||    :| |  | |      |:    ||
Db     531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy     572 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 617
           |  | |      :||      |  | |            :| |   ||     | ||
Db     586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy     618 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 661
           :| |  | |       :||      |  | |           :| | :  |  |      | |
Db     645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy     662 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 707
           |  :       :  :||      |  |    : ::|     | :  | |
Db     703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy     708 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 763
            :|   :||       |:  | |  | :|  ||     |:    :             |||
Db     763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy     764 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 823
```

```
;      STATE:  Virginia
;      COUNTRY:  United States
;      ZIP:  22313-1404
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/08/479,537A
;      FILING DATE:  07-JUN-1995
;      CLASSIFICATION:  514
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  FR 90/13101
;      FILING DATE:  23-OCT-1990
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  WO PCT/FR91/00835
;      FILING DATE:  23-OCT-1991
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/039,320
;      FILING DATE:  04-APR-1993
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/403,576
;      FILING DATE:  14-MAR-1995
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Teskin, Robin L.
;      REGISTRATION NUMBER:  35,030
;      REFERENCE/DOCKET NUMBER:  017753-025
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (703) 836-6620
;      TELEFAX:  (703) 836-2021
;  INFORMATION FOR SEQ ID NO:  2:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  2035 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  peptide
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  128..1899
;      OTHER INFORMATION:  /note= "The amino acids spanning
;      OTHER INFORMATION:  128 to 1899 constitute a repeated region wherein the repeat i
;      OTHER INFORMATION:  20 amino acids, 17 of which are fixed. The number of such
;      OTHER INFORMATION:  repeats varies from 1 to 40."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  134
;      OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;      OTHER INFORMATION:  Xaa Xaa which is the codon for Pro or Ala wherein Pro = CCT,
;      OTHER INFORMATION:  CCC, CCA, or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  144
;      OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;      OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC, ACA
;      OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  147
;      OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  1..21
;      OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;      OTHER INFORMATION:  21 amino acid precursor sequence."
US-08-479-537A-2

  Query Match             8.5%;  Score 423.5;  DB 2;  Length 2035;
  Best Local Similarity  24.4%;  Pred. No. 3.9e-19;

  Matches  235;  Conservative   63;  Mismatches  461;  Indels  203;  Gaps   48;

Qy      75 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 133
           | |||           | :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db      16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy     134 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 176
           |    :||:|    |    ||                | |      ||||     |   |||
Db      66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy     177 TTPKEPAPTTT-------KSAPTTPKEP------------APTTTKEPAPTTPKEPAPTT 217
             :||| :|          |||     |             ||      |  | |     |:
Db     123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy     218 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 264
             :  | | :|         |||     |  | |          ||      |  | |     |:
Db     181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy     265 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 311
             :  | | :  |         ||      |  |||         ||      |  | |     |:
Db     241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy     312 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 360
             :  |  |   ||       |||      |  |         |||     |  |
Db     301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy     361 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 412
            ||     | |        ||      |  | |         ||      |  |||     |
Db     360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy     413 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 468
           :  :  |    | |     |:    ||      | | |     |:  :  |     | |
Db     420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy     469 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 523
            | |    :||:     |  | |      |:  :  |  |  |||     | ||     |
Db     472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy     524 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 571
           | |           :| |    ||      ||    :| |  | |     |:    ||
Db     531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy     572 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 617
             | | |       :||      |  | |           :| |    ||     | ||
Db     586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy     618 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 661
           :| |  | |       :||      |  | |           :| | :  | |       | |
Db     645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy     662 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 707
             | :       :  :||       |  |     : ::|     | :  | |
Db     703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy     708 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 763
            :|   :||       |:  | |  | :|  ||      |:    :                |||
Db     763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy     764 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 823
                |     |  |  ||       :  :    |  ||      |::   | |
Db     821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy     824 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPT-SRIAEAMLQTTTRPNQ 882
            : ||     |  |  |     | |    :||  |:  |     |      |: |:
Db     869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPXPGSTAPXAHGVTSAPDX 924

Qy     883 TP 884
           |
Db     925 RP 926


RESULT  14
US-09-083-116-2
```

```
;       FILING DATE:
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  WO PCT/FR91/00835
;       FILING DATE:  23-OCT-1991
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  US 08/039,320
;       FILING DATE:  04-APR-1993
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  US 08/403,576
;       FILING DATE:  14-MAR-1995
;     ATTORNEY/AGENT INFORMATION:
;       NAME:  Teskin, Robin L.
;       REGISTRATION NUMBER:  35,030
;       REFERENCE/DOCKET NUMBER:  017753-025
;     TELECOMMUNICATION INFORMATION:
;       TELEPHONE:  (703) 836-6620
;       TELEFAX:  (703) 836-2021
;   INFORMATION FOR SEQ ID NO:  5:
;     SEQUENCE CHARACTERISTICS:
;       LENGTH:  1867 amino acids
;       TYPE:  amino acid
;       STRANDEDNESS:  single
;       TOPOLOGY:  linear
;     MOLECULE TYPE:  peptide
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  128..1727
;       OTHER INFORMATION:  /note= "The amino acids spanning
;       OTHER INFORMATION:  128 to 1727 constitute a repeated region wherein the repeat i
;       OTHER INFORMATION:  20 amino acids, 17 of which are fixed.  The number of such
;       OTHER INFORMATION:  repeats varies from 1 to 40."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  134
;       OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;       OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;       OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  144
;       OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;       OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC, ACA
;       OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  147
;       OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;       OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;       OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  1..21
;       OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;       OTHER INFORMATION:  21 amino acid precursor sequence."
US-09-083-116-5


  Query Match             8.5%;  Score 423.5;  DB 4;  Length 1867;
  Best Local Similarity   24.4%;  Pred. No. 3.5e-19;
  Matches  235;  Conservative   63;  Mismatches  461;  Indels  203;  Gaps   48;

Qy      75 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 133
           | |||          | :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db      16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy     134 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 176
          |    :||:|    |    ||               | |     ||||      |  |||
Db      66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy     177 TTPKEPAPTTT-------KSAPTTPKEP-----------APTTTKEPAPTTPKEPAPTT 217
           :||| :|        |||      |                ||      | | |     |:
Db     123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy     218 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 264
             :  | | :|         |||      |  | |        ||      |  | |     |:
Db     181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy     265 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 311
             :  | | :  |        ||      |  |||        ||      |  | |     |:
Db     241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy     312 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 360
             :  |  |   ||       |||      |  |          |||      |  |
Db     301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy     361 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 412
            ||      |  | |        ||      |  | |        ||      |  |||     |
Db     360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy     413 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 468
           :  :  |    | |       |:    ||      |  | |     |:  :  |     |  |
Db     420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy     469 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 523
           | |     :||:     |  | |      |:  :  |  |  |||      | ||     |
Db     472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy     524 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 571
           | |           :| |   ||       ||     :| |  | |      |:     ||
Db     531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy     572 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 617
           |  | |       :||     |  | |                :| |   ||     | ||
Db     586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy     618 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 661
           :| |   | |       :||      |  | |          :| | :  | |      | |
Db     645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy     662 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 707
             | :       :  :||      |  |    : ::|     | :  | |
Db     703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy     708 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 763
            :|   :||       |:  | |  | :|  ||      |:    :           |||
Db     763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy     764 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 823
                 |       |  | ||      :  :     |  ||      |::    | |
Db     821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy     824 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPT-SRIAEAMLQTTTRPNQ 882
           : ||      |  |    |       | |    :||  |:  |      |      |: |:
Db     869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPXPGSTAPXAHGVTSAPDX 924

Qy     883 TP 884
           |
Db     925 RP 926


RESULT  13
US-08-479-537A-2
; Sequence 2, Application US/08479537A
; Patent No. 5861381
;  GENERAL INFORMATION:
;    APPLICANT:  CHAMBON, Pierre
;    APPLICANT:  KIENY, Marie-Paule
;    APPLICANT:  LATHE, Richard
;    APPLICANT:  HAREUVENI, Mara
;    TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;    TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
;    NUMBER OF SEQUENCES:  5
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
;      STREET:  P.O. Box 1404
;      CITY:  Alexandria
```

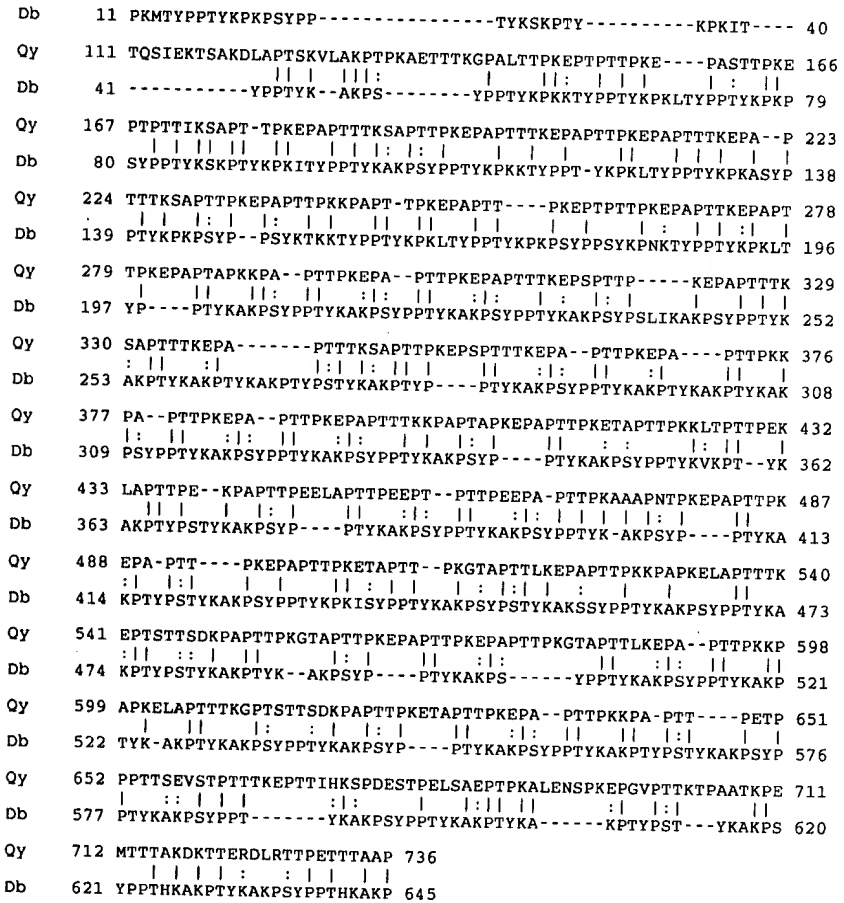
```
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (703) 836-6620
;      TELEFAX:  (703) 836-2021
;  INFORMATION FOR SEQ ID NO:  5:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  1867 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  peptide
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  128..1727
;      OTHER INFORMATION:  /note= "The amino acids spanning
;      OTHER INFORMATION:  128 to 1727 constitute a repeated region wherein the repeat i
;      OTHER INFORMATION:  20 amino acids, 17 of which are fixed.  The number of such
;      OTHER INFORMATION:  repeats varies from 1 to 40."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  134
;      OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  144
;      OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;      OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC, ACA
;      OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  147
;      OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  1..21
;      OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;      OTHER INFORMATION:  21 amino acid precursor sequence."
US-08-479-537A-5


  Query Match             8.5%;  Score 423.5;  DB 2;  Length 1867;
  Best Local Similarity   24.4%;  Pred. No. 3.5e-19;
  Matches  235;  Conservative   63;  Mismatches  461;  Indels  203;  Gaps   48;

Qy      75 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 133
           | |||           | :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db      16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy     134 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 176
           |    :||:|    |    ||             | |     ||||     |   |||
Db      66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy     177 TTPKEPAPTTT-------KSAPTTPKEP------------APTTTKEPAPTTPKEPAPTT 217
               :||| :|        |||      |             ||      |  | |     |:
Db     123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy     218 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 264
             :  | | :|        |||      |  | |         ||      |  | |     |:
Db     181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy     265 PKE--PAPTTKEP--------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 311
             :  | | :  |        ||      |  |||         ||      |  | |     |:
Db     241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy     312 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 360
             :  |  |   ||      |||      |  |         |||      |  |
Db     301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy     361 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 412
             ||      |  | |        ||      |  | |         ||      |  |||     |

Db     360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy     413 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 468
           :  : |     | |      |:     ||      |  | |      |:  :  |     |  |
Db     420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy     469 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 523
             | |     :||:     |  | |      |:  :  |  |  |||      | ||      |
Db     472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy     524 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 571
             | |           :| |    ||       ||     :| |  | |      |:    ||
Db     531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy     572 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 617
             |  | |       :||      |  | |           :| |   ||     | ||
Db     586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy     618 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 661
           :| |  | |       :||      |  | |           :| | :  | |       | |
Db     645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy     662 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 707
               | :       :  :||      |  |    : ::|     | :  | |
Db     703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy     708 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 763
            :|   :||       |:  | |  | :|  ||      |:    :             |||
Db     763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy     764 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 823
                     |       |   |  ||      :  :    |  ||      |::   | |
Db     821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy     824 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPT-SRIAEAMLQTTTRPNQ 882
             : ||      |  |   |       | |    :|| |:  |      |      |: |:
Db     869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPXPGSTAPXAHGVTSAPDX 924

Qy     883 TP 884
           |
Db     925 RP 926


RESULT  12
US-09-083-116-5
; Sequence 5, Application US/09083116
; Patent No. 6203795
;  GENERAL INFORMATION:
;    APPLICANT:  CHAMBON, Pierre
;    APPLICANT:  KIENY, Marie-Paule
;    APPLICANT:  LATHE, Richard
;    APPLICANT:  HAREUVENI, Mara
;    TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;    TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
;    NUMBER OF SEQUENCES:  5
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
;      STREET:  P.O. Box 1404
;      CITY:  Alexandria
;      STATE:  Virginia
;      COUNTRY:  United States
;      ZIP:  22313-1404
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/09/083,116
;      FILING DATE:
;      CLASSIFICATION:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  08/479,537
```

```
Qy    798 IMNKP--EETAKPKD-------RATNSKATTPKPQKPTKAPKKPTSTKKPKTMP------ 842
          :   | | |  |||        :|  :   |   ||         : ||   |
Db   1024 VYGLPVGEITGLPKDPGSDIPFNSTTGELVDPSTGKPINNSTAGIVSGKPGLPPIEDENG 1083

Qy    843 --------------RVRKPKTTPTPRKMTS--TMPE---------------------- 862
                       ::  |:|  |    ||  | |:
Db   1084 NLFDPSTNLPIDGNNQLVNPETNSTVSGSTSGTTKPKPGIPVNGGGVVPDEEAKDQADKG 1143

Qy    863 -----LNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMP 917
               : ||: | :  :  |   | | |     :||  |      |  | |  |    |
Db   1144 KDGLIVPPTNSINKDPVTNTQYSNTTGNI----INP--ETGKVIPGSLPGSLNYPSFNTP 1197

Qy    918 EVTPD--------MDYLPRVPNQGIIINP 938
          : | :        :  ||  |: | ||:|
Db   1198 QQTDEITGKPVDTVTGLPYDPSTGEIIDP 1226


RESULT  10
US-09-219-849-5
; Sequence 5, Application US/09219849
; Patent No. 6150081
; GENERAL INFORMATION:
;  APPLICANT:  VAN HEERDE, GEORGE V.
;  APPLICANT:  VAN RIJN, ALEXIS C.
;  APPLICANT:  BOUWSTRA, JAN B.
;  APPLICANT:  DE WOLF, FREDERIK A.
;  APPLICANT:  MOOBROEK, ANDREAS
;  APPLICANT:  WERTEN, MARC W.T.
;  APPLICANT:  WIND, RICHELE D.
;  APPLICANT:  VAN DEN BOSCH, TANJA J.
;  TITLE OF INVENTION:  SILVER HALIDE EMULSIONS WITH RECOMBINANT COLLAGEN
;  TITLE OF INVENTION:  SUITABLE FOR PHOTOGRAPHIC APPLICATION AND ALSO THE
;  TITLE OF INVENTION:  PREPARATION THEREOF
;  FILE REFERENCE:  2728-2
;  CURRENT APPLICATION NUMBER: US/09/219,849
;  CURRENT FILING DATE:  1998-12-23
;  NUMBER OF SEQ ID NOS:  50
;  SOFTWARE:  PatentIn Ver. 2.1
; SEQ ID NO 5
;  LENGTH: 960
;  TYPE: PRT
;  ORGANISM: Artificial Sequence
;  FEATURE:
;  OTHER INFORMATION: Description of Artificial Sequence: Illustrative
;  OTHER INFORMATION: amino acid sequence
US-09-219-849-5


  Query Match             9.0%;  Score 452;  DB 4;  Length 960;
  Best Local Similarity   25.6%;  Pred. No. 2.8e-21;
  Matches  173;  Conservative   60;  Mismatches  309;  Indels  134;  Gaps   41;

Qy    132 PTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAP--TTTKS 189
          | |     :: |    |  | | |  | | |  |  |      : |   ::| |     :
Db     72 PGPAGPPGSRDPG--PPGAPGPAGP--PGSRDPGPPGAPG-PAGPPGSRDPGPPGAPGPA 126

Qy    190 APTTPKEPAPTTTKEPA-PTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTP---KKPA 245
          |   ::| |     || |   ::| |      ||       |  |  | |  |   : |
Db    127 GPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPG 186

Qy    246 PTTPKEPAPTTP---KEPTPTTPKEPA--PTTKEPAPTTPKEPA--------------- 284
          |  | | |  |   ::| |     ||  | :::| |     ||
Db    187 P--PGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAHGPAGPKGAHGPAGPKGAHG 244

Qy    285 ---------PTAPK-KPAPTTP---KEPAPTTPKEPA-PTTTKEPSPTTPKEPAPTTTKS 330
                  |  ||  | | |   ::| |     || |  :::| | |  | | |     :
Db    245 PAGPKGAHGPAGPKGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP--PGAPGP----A 298

Qy    331 APTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTP---KKPAPTTPKEPAP 387
          |  :::| |      | |    |  :::| | |  | | |   : | |  |  | |
Db    299 GPPGSRDPGP------PGAPGPAGPPGSRDPGP--PGAPGPAGPPGSRDPGP--PGAPGP 348

Qy    388 TTP---KEPAPTTTKKPA-------PTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTT 437
            |   ::| |      ||       |  | | | |  :    |    |  |    |
Db    349 AGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP----PGAPGPAGPPG 404

Qy    438 PEKPAPT-TPEELAPTTPEEPTPTTPEEPAPTTPKAAA-PNTPKEPAPTTP---KEPAPT 492
          | |   |    |    :| | |  | | | |  :  |  | | |  |   ::| |
Db    405 SRDPGPPGAPGPAGPPGSRDPGP--PGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPP 462

Qy    493 TPKEPA-------PTTPKET-APTTPKGT--------APTTLKEPAPTTPKKP-APKELA 535
            ||       |  ||     |  |||                ||     |    |  | ||
Db    463 GAHGPAGPKGAHGPAGPKGAHGPAGPKGAHGPAGPKGAPGPAGPPGSRDPGPPGAPGPAG 522

Qy    536 PTTTKEPTSTTSDKPAPTTPKGTA-PTTPKEPAPTTP---KEPAPTTPKGTA-PTTLKEP 590
          |  :::|    :  | |  | |:  |  |  | |  |   ::| |    | | |   ::|
Db    523 PPGSRDPGPPGA--PGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDP 580

Qy    591 APTTPKKPAP------KELAPTTTKGPTST--TSDKPAPTTPKETAPTTPKEPAPTTPKK 642
            |  | | |       ::  |    ||      : |  |  |  |    |  ::| |  |
Db    581 GP--PGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP--PGA 636

Qy    643 PAPTTP----ETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALEN-----S 693
          | |  |    :  ||     : |  :::|     :|  : |  | :| |
Db    637 PGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP-PGAPGPAGPPGSRDPGPPGAPGPAGPPG 695

Qy    694 PKEPGVPTTKTPAATK 709
            ::|| |    ||  |
Db    696 SRDPGPPGAHGPAGPK 711


RESULT  11
US-08-479-537A-5
; Sequence 5, Application US/08479537A
; Patent No. 5861381
;  GENERAL INFORMATION:
;    APPLICANT:  CHAMBON, Pierre
;    APPLICANT:  KIENY, Marie-Paule
;    APPLICANT:  LATHE, Richard
;    APPLICANT:  HAREUVENI, Mara
;    TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;    TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
;    NUMBER OF SEQUENCES:  5
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
;      STREET:  P.O. Box 1404
;      CITY:  Alexandria
;      STATE:  Virginia
;      COUNTRY:  United States
;      ZIP:  22313-1404
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/08/479,537A
;      FILING DATE:  07-JUN-1995
;      CLASSIFICATION:  514
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  FR 90/13101
;      FILING DATE:  23-OCT-1990
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  WO PCT/FR91/00835
;      FILING DATE:  23-OCT-1991
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/039,320
;      FILING DATE:  04-APR-1993
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/403,576
;      FILING DATE:  14-MAR-1995
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Teskin, Robin L.
;      REGISTRATION NUMBER:  35,030
;      REFERENCE/DOCKET NUMBER:  017753-025
```

```
Db   1133 DEEAKDQADKGKDGLIVPPTNSINKDPVTNTQYSNTTGNIINPETGKVIPGSLPGSLNYP 1192

Qy    675 DESTPELSAEPTPKALE---NSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPET 731

Db   1193 SFNTPQQTDEITGKPVDTVTGLPYDPSTGEIIDPATKLPIPGSVAGDEILTEVLNITTDE 1252

Qy    732 TTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKK 791

Db   1253 VTGLP-IDLETGLPRDP------VSGLPQLPNGTLVD----------------PSNKK 1287

Qy    792 TI----------TTTEIMNKPEETAKPKD----------------------RATNSK 816

Db   1288 PIPGSHSGFINGTSGEQSHEKDPSTGKPLDPNTGLHPFDEDSGSLINPETGDKLQGSHSG 1347

Qy    817 ATTPKPQKP------------TKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTST---- 859

Db   1348 TFMPVPGKPQGENGGIMTPEQILEALNKLPTSNEVNISPR---PSSDAVPDRPTNTWWNK 1404

Qy    860 ----------MPELNPTSRIAEAMLQTTTRPNQT------PNSKL-------VEVNPKS 895

Db   1405 ISGQTYQVDGKKTILGSAASVIHTALGTPTQTDPTTGLPSDPSTGLPFIPGFNVLVDPQT 1464

Qy    896 EDAGGAEGETPHMLLRPHVFMPEVTPDMDY-LPRVPNQGIIINPM 939

Db   1465 GEQ--IKGSVPYVSL--YVKEKNIVTEAAYGLPVDPKTGFPIDPI 1505
```

```
RESULT   9
US-08-928-361B-6
; Sequence 6, Application US/08928361B
; Patent No. 6071518
;  GENERAL INFORMATION:
;    APPLICANT: Petersen, Carolyn
;    TITLE OF INVENTION:  PEPTIDES, POLYPEPTIDES, GLYCOPROTEINS,
;    TITLE OF INVENTION:  THEIR FUNCTIONAL MUTANTS, VARIANTS, ANALOGS AND FRAGMENTS
;    TITLE OF INVENTION:  FOR TREATMENT AND DETECTION/DIAGNOSIS OF CRYPTOSPORIDIUM
;    TITLE OF INVENTION:  SPECIES INFECTIONS
;    NUMBER OF SEQUENCES:  30
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  PETERS, VERNY, JONES & BIKSA
;      STREET:  385 Sherman Avenue, Suite 6
;      CITY:  Palo Alto
;      STATE:  CA
;      COUNTRY:  USA
;      ZIP:  94306-1840
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/08/928,361B
;      FILING DATE:  12-SEP-1997
;      CLASSIFICATION:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 60/026,062
;      FILING DATE:  13-SEP-1996
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Verny, Hana
;      REGISTRATION NUMBER:  30,518
;      REFERENCE/DOCKET NUMBER:  480.76-1(HV)
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  650-324-1677
;      TELEFAX:  650-324-1678
;  INFORMATION FOR SEQ ID NO:  6:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  1721 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
US-08-928-361B-6
```

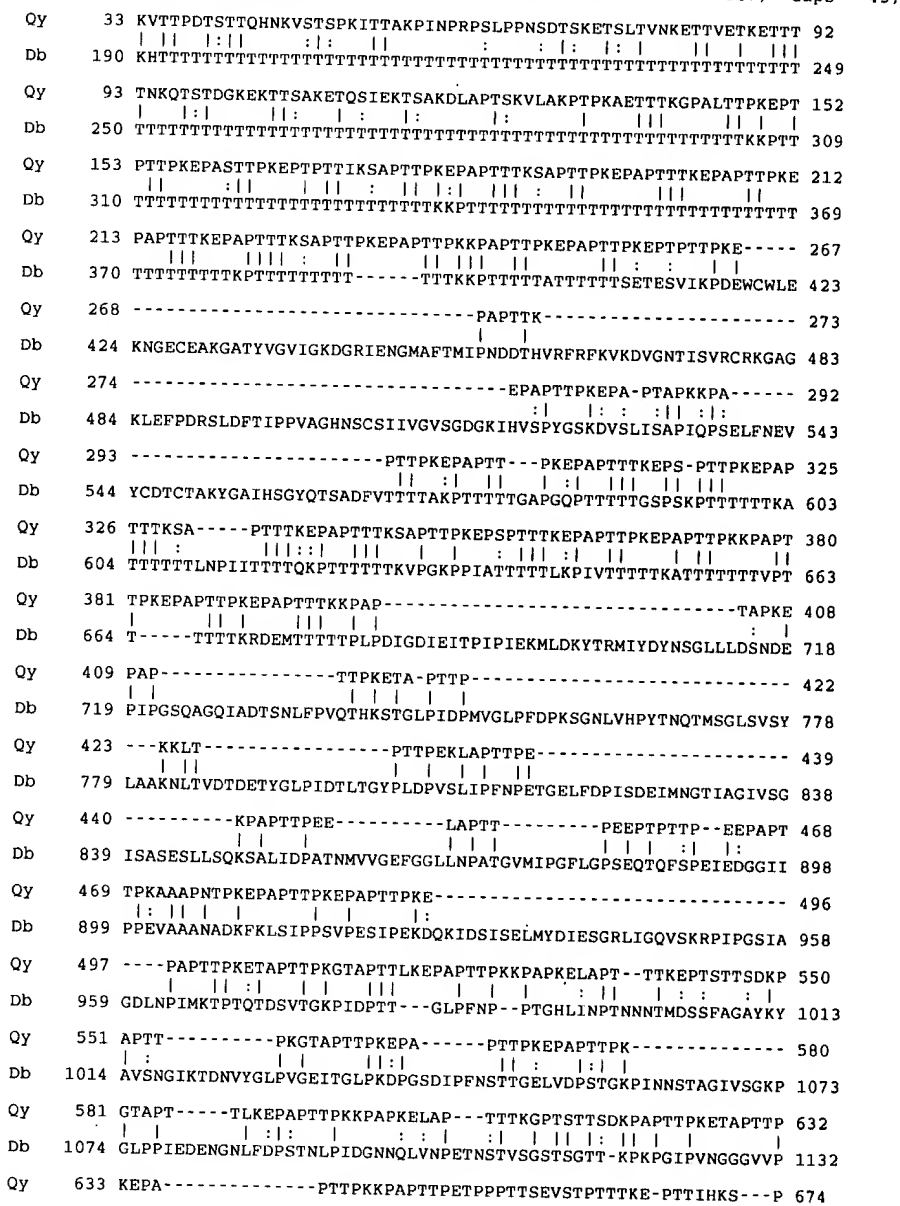
```
Query Match           9.2%;  Score 463.5;  DB 3;  Length 1721;
Best Local Similarity  22.1%;  Pred. No. 9.7e-22;
Matches  271;  Conservative   78;  Mismatches  467;  Indels  413;  Gaps   39;

Qy      5 KKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPRPS 64

Db    116 RSNETK---TTEPSANTYAGVYRSN-ETKTTEPSANT---NFLLVDPKI----------N 158

Qy     65 LPPNSDTSKE--------------TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAK 109

Db    159 APCNSENSFEQGQIFDMGSKVYIPYTKCVGVKHTTTTTTTTTTTTTTTT---------- 208

Qy    110 ETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTP 169

Db    209 ----------------------------TTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 240

Qy    170 TTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSA 229

Db    241 TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 300

Qy    230 PTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPK 289

Db    301 TTTTKKPTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTKKPTTTTTTTTTTT---------- 350

Qy    290 KPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTT 349

Db    351 ----TTTTTTTTTTTTTTTTTTTTTTTTTTTKPTTTTTTTTTTTTTKKPTTTTTATTTTT 406

Qy    350 PKE-------------------------------------------------PSPTT------- 357

Db    407 TSETESVIKPDEWCWLEKNGECEAKGATYVGVIGKDGRIENGMAFTMIPNDDTHVRFRFK 466

Qy    358 ----------------------------TKEP--------------------APTTPKE 368

Db    467 VKDVGNTISVRCRKGAGKLEFPDRSLDFTIPPVAGHNSCSIIVGVSGDGKIHVSPYGSKD 526

Qy    369 PAPTTPKKPAPTTPKE-----------------PAPTTPKEPAPTTTKKPAPTAPKEPA 410

Db    527 VSLIS----APIQPSELFNEVYCDTCTAKYGAIHSGYQTSADFVTTTTAKPTTTT--TGA 580

Qy    411 PTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTP 470

Db    581 PGQPTTTTTGSPSKPTTTTTTKATTTT------TTLNPIITTTTQKPTTTT------TT- 627

Qy    471 KAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPA 530

Db    628 --KVPGKPPIATTTTTLKPIVTTTTTKATTTTTTTVPTT-----TTTTKRDEMTTTTTPL 680

Qy    531 PK----ELAPTTTKEPTS--------------TTSDKPAPTTPKG-TAPTTPKEPAPTT 570

Db    681 PDIGDIEITPIPIEKMLDKYTRMIYDYNSGLLLDSNDEPIPGSQAGQIADTSNLFPVQTH 740

Qy    571 PKEPAPTTPKGTAPTTLKE---PAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKET 627

Db    741 KSTGLPIDPMVGLPFDPKSGNLVHPYTNQTMSGLSVSYLAAKNLTVDTDETYG--LPIDT 798

Qy    628 APTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPT--------TTKEPTTIHKS------ 673

Db    799 LTGYPLDPVSLIPFN--PETGELFDPISDEIMNGTIAGIVSGISASESLLSQKSALIDPA 856

Qy    674 ------------------------PDEST-------------PELSAEPTPKALENSPK 695

Db    857 TNMVVGEFGGLLNPATGVMIPGFLGPSEQTQFSPEIEDGGIIPPEVAAANADKFKLSIP- 915

Qy    696 EPGVPTTKTPAATKPEMTTTAKDKTTER----------------DLRTTPETTTAAPKM 738

Db    916 -PSVP------ESIPE-----KDQKIDSISELMYDIESGRLIGQVSKRPIPGSIAGDLNP 963

Qy    739 TKETATTTEKTTESKITATT-TQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTE 797

Db    964 IMKTPTQTDSVTGKPIDPTTGLPFNPPTGHLINPTNNNTMDSSFAGAYKYAVSNGIKTDN 1023
```



RESULT 8
US-08-700-651-5
; Sequence 5, Application US/08700651B
; Patent No. 6015882
; GENERAL INFORMATION:
; APPLICANT: PETERSEN, CAROLYN
; APPLICANT: LEECH, JAMES
; APPLICANT: NELSON, RICHARD, C.
; APPLICANT: GUT, JIRI
; TITLE OF INVENTION: VACCINES, ANTIBODIES, PROTEINS, GLYCOPROTEINS, DNAS AND RNAS
; TITLE OF INVENTION: FOR PROPHYLAXIS AND TREATMENT OF Cryptosporidium parvum
; TITLE OF INVENTION: INFECTIONS
; FILE REFERENCE: 480.19-4(HV)
; CURRENT APPLICATION NUMBER: US/08/700,651B
; CURRENT FILING DATE: 1997-08-14
; EARLIER APPLICATION NUMBER: 08/415,751
; EARLIER FILING DATE: 1995-04-03
; NUMBER OF SEQ ID NOS: 15
; SOFTWARE: PatentIn Ver. 2.0
; SEQ ID NO 5
; LENGTH: 1721
; TYPE: PRT

; ORGANISM: Cryptosporidium parvum
US-08-700-651-5

Query Match 9.3%; Score 466.5; DB 3; Length 1721;
Best Local Similarity 20.9%; Pred. No. 6.3e-22;
Matches 285; Conservative 100; Mismatches 473; Indels 507; Gaps 45;



```
Qy     335 TKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA-PTTPKKP-APTTPKEPAPTTPKE 392
           :| |    : |: | :|:|       |  | ||: |  |  | ||: | |  |: | |
Db     444 PNKPNP----NEPSNPNKPNPN-----EPLNPNEPSNPNEPSNPNAPSNPNE--PSNPNE 492

Qy     393 PAPTTTKKPAPTAPKEP-APTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAP 451
           |:        |  |  |  |: | |  |: |||                    |: | |  |
Db     493 PS-------NPNEPSNPNEPSNPNE--PSNPKK-----------------PSNPNE--P 523

Qy     452 TTPEEPTPTTPEEPAPTTPKAAAPNTPKEP-APTTPKEPAPTTPKEPAPTTPKETAPTTP 510
           : | |  |  | ||       : || |  |  |: |:|  |: |||  |: | |
Db     524 SNPNE--PLNPNEP-------SNPNEPSNPNEPSNPEE--PSNPKE--PSNPNE------ 564

Qy     511 KGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTT 570
                   |: |::| |:|  |:  |||    |:  |  |:  |  | |  :
Db     565 -----------PSNPEEPNPEE--PSNPKEP----SNPEEPINPEELNPKEPSNPEESN 606

Qy     571 PKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPT 630
           ||||                 |  : ||:|    : |  : |     :  |       |
Db     607 PKEPI------------NPEESNPKEP----INPEDNENPLIIQDEPIEPRNDSNVIPI 649

Qy     631 TPKEP------APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAE 684
           |  |          |: |: :  | | |  :  ::  | |    |   |:|  |
Db     650 LPIIPQKGNNIPSNLPENPSDSEVEYPRPNDNGENSNNTMKSKKNI---PNEPIPSPGDN 706

Qy     685 PTPKALENSPK-------------------EPGVPTTK 703
           |     |  ||                  || :|  :
Db     707 PYKGHEERIPKPHRSNDDYVYDNNVNKNNKDEPEIPNNE 745
```

```
RESULT   6
PCT-US92-00018-2
; Sequence 2, Application PC/TUS9200018
;  GENERAL INFORMATION:
;     APPLICANT:  Hoffman, Stephen L.
;     APPLICANT:  Charoenvit, Yupin
;     APPLICANT:  Hedstrom, Richard
;     APPLICANT:  Khusmith, Srisin
;     APPLICANT:  Rogers IV, William O.
;     TITLE OF INVENTION:  Protective malaria sporozoite surface protein
;     TITLE OF INVENTION:  immunogen and gene encoding
;     NUMBER OF SEQUENCES:  2
;     CORRESPONDENCE ADDRESS:
;       ADDRESSEE:  A. David Spevack
;       STREET:  NMRDC Building 1 T-12 National Naval
;       STREET:  Medical Center
;       CITY:  Bethesda
;       STATE:  MD
;       COUNTRY:  USA
;       ZIP:  20814-5044
;     COMPUTER READABLE FORM:
;       MEDIUM TYPE:  Floppy disk
;       COMPUTER:  IBM PC compatible
;       OPERATING SYSTEM:  PC-DOS/MS-DOS
;       SOFTWARE:  PatentIn Release #1.24
;     CURRENT APPLICATION DATA:
;       APPLICATION NUMBER:  PCT/US92/00018
;       FILING DATE:  19920103
;       CLASSIFICATION:  424
;     ATTORNEY/AGENT INFORMATION:
;       NAME:  Spevack, Avram D.
;     TELECOMMUNICATION INFORMATION:
;       TELEPHONE:  (301) 295-6759
;       TELEFAX:  (301) 295-4033
;   INFORMATION FOR SEQ ID NO:  2:
;     SEQUENCE CHARACTERISTICS:
;       LENGTH:  826 amino acids
;       TYPE:  AMINO ACID
;       TOPOLOGY:  linear
;     MOLECULE TYPE:  protein
PCT-US92-00018-2
```

```
  Query Match             9.7%;  Score 488.5;  DB 5;  Length 826;
  Best Local Similarity   26.4%;  Pred. No. 1.2e-23;
  Matches  153;  Conservative   54;  Mismatches  223;  Indels  149;  Gaps   34;

Qy     168 TPTTIKSAPTTPKEPA-PTTTKSAPTTPKEPA-PTTTKEP----APTTPKEP-APTTTKE 220
           ||  ::  |  |  |  |       |: |:||  |     :|      |  |  |  |
Db     273 TPCKVRDCPQIPIPPVIPNKIPEKPSNPEEPVNPNDPNDPNNPNNPNNPNNPNNPNNN 332

Qy     221 PAPTTTKSAPTTPKEP-APTTPKKP-APTTPKEP-APTTPKEP-TPTTPKEPAPTTKEPA 276
           |      : |  |  |  |  |  |  |  |  |  |  |  |  |  | :|:     |
Db     333 PNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPDPSNPNNHPK 392

Qy     277 PTTPKEPAPTAPK--KPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTT 334
           ||   |  ||  || |  |    |: | :| |       |: | :| |    : |:
Db     393 RRNPKRRNPNKPKPNKPNPNKPNPNEPSNPNKPNPN-----EPSNPNKPNP----NEPSN 443

Qy     335 TKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA-PTTPKKP-APTTPKEPAPTTPKE 392
           :| |    : |: | :|:|       |  | ||: |  |  | ||: | |  |: | |
Db     444 PNKPNP----NEPSNPNKPNPN-----EPLNPNEPSNPNEPSNPNAPSNPNE--PSNPNE 492

Qy     393 PAPTTTKKPAPTAPKEP-APTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAP 451
           |:        |  |  |  |: | |  |: |||                    |: | |  |
Db     493 PS-------NPNEPSNPNEPSNPNE--PSNPKK-----------------PSNPNE--P 523

Qy     452 TTPEEPTPTTPEEPAPTTPKAAAPNTPKEP-APTTPKEPAPTTPKEPAPTTPKETAPTTP 510
           : | |  |  | ||       : || |  |  |: |:|  |: |||  |: | |
Db     524 SNPNE--PLNPNEP-------SNPNEPSNPNEPSNPEE--PSNPKE--PSNPNE------ 564

Qy     511 KGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTT 570
                   |: |::| |:|  |:  |||    |:  |  |:  |  | |  :
Db     565 -----------PSNPEEPNPEE--PSNPKEP----SNPEEPINPEELNPKEPSNPEESN 606

Qy     571 PKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPT 630
           ||||                 |  : ||:|    : |  : |     :  |       |
Db     607 PKEPI------------NPEESNPKEP----INPEDNENPLIIQDEPIEPRNDSNVIPI 649

Qy     631 TPKEP------APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAE 684
           |  |          |: |: :  | | |  :  ::  | |    |   |:|  |
Db     650 LPIIPQKGNNIPSNLPENPSDSEVEYPRPNDNGENSNNTMKSKKNI---PNEPIPSPGDN 706

Qy     685 PTPKALENSPK-------------------EPGVPTTK 703
           |     |  ||                  || :|  :
Db     707 PYKGHEERIPKPHRSNDDYVYDNNVNKNNKDEPEIPNNE 745
```

```
RESULT   7
5202236-13
;Patent No. 5202236
;    APPLICANT: MAUGH, KATHY J.;ANDERSON, DAVID M.;STRAUSBERG,
;SUSAN L.;MCCANDLISS, RUSS;WEI, TENA;FILPULA, DAVID
;    TITLE OF INVENTION: METHOD OF PRODUCING BIOADHESIVE
;PROTEIN
;    NUMBER OF SEQUENCES: 39
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER: US/07/528,762
;      FILING DATE: 25-MAY-1990
;      APPLICATION NUMBER: 82,456
;      FILING DATE: 07-AUG-1987
;      APPLICATION NUMBER: 933,945
;      FILING DATE: 24-NOV-1986
;      APPLICATION NUMBER: 650,128
;      FILING DATE: 13-SEP-1984
;SEQ ID NO:13:
;      LENGTH: 652
5202236-13
```

```
  Query Match             9.5%;  Score 476.5;  DB 6;  Length 652;
  Best Local Similarity   29.7%;  Pred. No. 5.3e-23;
  Matches  221;  Conservative   71;  Mismatches  284;  Indels  169;  Gaps   45;

Qy      51 PKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKE 110
           ||:|      |:|| ||                | ::| |           | | |
```

```
  Matches  273;  Conservative   86;  Mismatches  459;  Indels  432;  Gaps   41;

Qy      33 KVTTPDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTT 92
Db     307 KHTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 366

Qy      93 TNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPT 152
Db     367 TTTTTTTTKKPTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTKKPTT 426

Qy     153 PTTPKEPASTTPKEPTPTTIKSAPTTPKEP--APTTTKSAPTTPKEPAPTTTKEPAPTTP 210
Db     427 TTTTTTTTTTTTTTTTTTTTTTTTTTKKPTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 486

Qy     211 KEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKE---------- 259
Db     487 TTTATTTTKKPTTTTTTTTTTTTKKPTTTTTATTTTTTSETESVIKPDEWCWLEKNGECE 546

Qy     260 ------------------------PTPTT--------------------------PK 266
Db     547 AKGATYVGVIGKDGRIENGMAFTMIPNDDTHVRFRFKVKDVGNTISVRCGKGAGKLEFPD 606

Qy     267 EPAPTTKEP--------------------APTTPKEPA-PTAPKKP---------- 291
Db     607 RSLDFTIPPVAGHNSCSIIVGVSGGGKIHVSPYGSKDVSLISAPIQPCELFNEVYCDTCT 666

Qy     292 ----------------APTTPKEPAPTT---PKEPAPTTTKEPS-PTTPKEPAPTTTKS- 330
Db     667 AKYGAIHSGYQTSADFVTTTTAKPTTTTTGAPGQPTTTTTGSPSKPTTTTTTKATTTTTI 726

Qy     331 ----APTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPA 386
Db     727 LNPIITTTTQKPTTTTTTKVPGKPPIATTTTTTLKPIVTTTTTKATTTTTTTVPTT----- 781

Qy     387 PTTPKEPAPTTTKKPAP------------------------------TAPKEPAP--- 411
Db     782 TTTTKRDEMTTTTTPLPDIGDIEITPIPIEKMLDKYTRMIYDYNSGLLLDSNDEPIPGSQ 841

Qy     412 ------------TTPKETA-PTTP------------------------------KKL 425
Db     842 AGQIADTSNLFPVQTHKSTGLPIDPMVGLPFDPKSGNLVHPYTNQTMSGLSVSYLAAKNL 901

Qy     426 T----------------PTTPEKLAPTTPE---------------------------- 439
Db     902 TVDTDETYGLPIDTLTGYPLDPVSLIPFNPETGELFDPISDEIMNGTIAGIVSGISASES 961

Qy     440 ----KPAPTTPEE----------LAPTT---------PEEPTPTTP--EEPAPTTPKAAA 474
Db     962 LLSQKSAPIDPATNMVVGEFGGLLNPATGVMIPGSLGPSEQTPFSPEIEDGGIIPPEVAA 1021

Qy     475 PNTPKEPAPTTPKEPAPTTPKE---------------------------------PA 498
Db    1022 ANADKFKLSIPPSVPESIPEKDQKIDSISELMYDIESGRLIGQVSKRPIPGSIAGDLNPI 1081

Qy     499 PTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGT 558
Db    1082 MKTPTQTDSVTGKPIDPTT---GLPFNP--PTGHLINPTNNNTMDSSFAGAYKYAVSNGI 1136

Qy     559 APTT----PKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTP----------KKPAPK-- 601
Db    1137 KTDNVYGLPVDEITGLPKDPVSDIPFNSTTGELVDPSTGKPINNYTAGIVSGKRGLPPIE 1196

Qy     602 ----ELAPTTTKGP----------TSTT-------SDKPAPTTPKETAPTTPKEPA--- 636
Db    1197 DENGNLFDPSTKLPIDGNNQLVNPETNSTVSGSTSGSTKPKPGIPVNGGGVVPDEEAKDQ 1256

Qy     637 ----------PTTPKKPAPTTPETPPPTTSEVSTPTTTKE-PTTIHKS---PDESTPEL 681
Db    1257 ADKGKDGLIVPPTNSINKDPVTNTQYSNTTGNIINPETGKVIPGSLPGSLNYPSFNTPQQ 1316

Qy     682 SAEPTPKALE---NSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKM 738
Db    1317 TDEITGKPVDTVTGLPYDPSTGEIIDPATKLPIPGSVAGDEILTEVLNITTDEVTGLP-I 1375

Qy     739 TKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTE- 797
Db    1376 DLETGLPRDP------VSGLPQLPNGTLVD----------------PSNKKPIPGSHS 1411

Qy     798 --IMNKPEETAKPKDRATNSKATTPKPQKP-TKAP--KKPTSTKKPKTMPRVRKPKT--- 849
Db    1412 GFINGTSGEQSHEKDPSTG------KPLDPNTGLPFDEDSGSLINPETGDKLQGSHSGTF 1465

Qy     850 TPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPK-SEDA 898
Db    1466 MPVPGKPQGENGGIMTPEQILEAL-------NKLPTSNEVNISPRPSSDA 1508
```

RESULT 5
US-07-638-431-2
; Sequence 2, Application US/07638431
; Patent No. 5198535
; GENERAL INFORMATION:
; APPLICANT: Hoffman, Stephen L.
; APPLICANT: Charoenvit, Yupin
; APPLICANT: Hedstrom, Richard
; APPLICANT: Khusmith, Srisin
; APPLICANT: Rogers IV, William O.
; TITLE OF INVENTION: Protective malaria sporozoite surface protein
; TITLE OF INVENTION: immunogen and gene
; NUMBER OF SEQUENCES: 2
; CORRESPONDENCE ADDRESS:
; ADDRESSEE: A. David Spevack
; STREET: NMRDC Building 1 T-12 National Naval
; STREET: Medical Center
; CITY: Bethesda
; STATE: MD
; COUNTRY: USA
; ZIP: 20814-5044
; COMPUTER READABLE FORM:
; MEDIUM TYPE: Floppy disk
; COMPUTER: IBM PC compatible
; OPERATING SYSTEM: PC-DOS/MS-DOS
; SOFTWARE: PatentIn Release #1.24
; CURRENT APPLICATION DATA:
; APPLICATION NUMBER: US/07/638,431
; FILING DATE: 19910110
; CLASSIFICATION: 424
; ATTORNEY/AGENT INFORMATION:
; NAME: Spevack, Avrom D.
; TELECOMMUNICATION INFORMATION:
; TELEPHONE: (301) 295-6759
; TELEFAX: (301) 295-4033
; INFORMATION FOR SEQ ID NO: 2:
; SEQUENCE CHARACTERISTICS:
; LENGTH: 826 amino acids
; TYPE: AMINO ACID
; TOPOLOGY: linear
; MOLECULE TYPE: protein
US-07-638-431-2

```
  Query Match             9.7%;  Score 488.5;  DB 1;  Length 826;
  Best Local Similarity   26.4%;  Pred. No. 1.2e-23;
  Matches  153;  Conservative   54;  Mismatches  223;  Indels  149;  Gaps   34;

Qy     168 TPTTIKSAPTTPKEPA-PTTTKSAPTTPKEPA-PTTTKEP----APTTPKEP-APTTTKE 220
Db     273 TPCKVRDCPQIPIPPVIPNKIPEKPSNPEEPVNPNDPNDPNNPNNPNNPNNPNNPNN 332

Qy     221 PAPTTTKSAPTTPKEP-APTTPKKP-APTTPKEP-APTTPKEP-TPTTPKEPAPTTKEPA 276
Db     333 PNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPDPSNPNNHPK 392

Qy     277 PTTPKEPAPTAPK--KPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTT 334
Db     393 RRNPKRRNPNKPKPNKPNPNKPNPNEPSNPNKPNPN-----EPSNPNKPNP----NEPSN 443
```

```
;  GENERAL INFORMATION:
;    APPLICANT:  Granados, Robert R
;    APPLICANT:  Wang, Ping
;    TITLE OF INVENTION:  A No. 6187558el Invertebrate Intestinal Mucin
;    TITLE OF INVENTION:  cDNA and Related Products and Methods
;    NUMBER OF SEQUENCES:  4
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  Brown, Pinnisi & Michaels, P.C.
;      STREET:  118 No. 6187558th Tioga
;      CITY:  Ithaca
;      STATE:  NY
;      COUNTRY:  USA
;      ZIP:  14850
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/09/103,429A
;      FILING DATE:  24-JUN-1998
;      CLASSIFICATION:  800
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Michaels, Christopher A
;      REGISTRATION NUMBER:  34,390
;      REFERENCE/DOCKET NUMBER:  BTI-39
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (607) 256-2000
;      TELEFAX:  (607) 256-3628
;  INFORMATION FOR SEQ ID NO:  3:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  786 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
;    HYPOTHETICAL:  NO
;    ANTI-SENSE:  NO
;    ORIGINAL SOURCE:
;      ORGANISM:  Trichoplusia ni
;      TISSUE TYPE:  peritrophic membrane
US-09-103-429A-3


  Query Match             9.9%;  Score 498.5;  DB 4;  Length 786;
  Best Local Similarity   29.1%;  Pred. No. 2.6e-24;
  Matches  200;  Conservative   25;  Mismatches  207;  Indels  255;  Gaps   32;

Qy    132 PTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAP 191
          | | || ||: || |      ||| : | :||           ||||  :   |||  ||
Db     93 PGPPAE-TTQAPATT----QAPTTTQAPTTTT----------QAPTTTTQ--ATTTTQAP 135

Qy    192 TTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKE 251
          || :  |||||:  | ||    |||||  : ||||  ||||  : ||||  :  ||||  :
Db    136 TTTQ--APTTTQ--ATTT---QAPTTTTQ-APTTTTQAPTTTTQ-APTTTQ--APTTTQ- 183

Qy    252 PAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTT-PKEPA-- 308
           |||| : ||  |      |    || |||    | |   || |||  ||||: |  |
Db    184 -APTTTQAPTTITQAATTPAATTPAATTPAATTPAA-TTPAATTPGVPAPTSAPVWPPIC 241

Qy    309 ----------------------------------------------PTTTKEPSPTTPK- 321
                                                         |   :  ||   :
Db    242 ELLPNGCPADFDIHLLIPHDKYCNLFYQCSNGYTFEQRCPEGLYFNPYVQRCDSPANVEC 301

Qy    322 ----EPAPTTT---------------------------------------------------- 328
              |||  |
Db    302 DGEISPAPPVTEGNEDEDIDIGDLLDNGCPANFEIDWLLPHGNRCDKYYQCVHGNLVERR 361

Qy    329 --------------------------KSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEP 361
                                       :|      | |  |    || | |  |     |
Db    362 CGAGTHFSFELQQCDHIELVGCTLPGGESEEVDVDEDA-CTGWYCPTEPIEWEPLPNGCP 420

Qy    362 APTT-----PKEP-----------------------APTTPKKPAPTT----------- 381
```

```
          |  :     | |                              :| |    :| |
Db    421 ADFSIDHLLPHESDCGQYLQCVHGQTIARPCPGNLHFSPATQSCESPVTAGCQVFECDSD 480

Qy    382 ---PKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTP 438
              |||     |||     ||||   |:   |  | | |   : |||     ||
Db    481 NQCTSTAAPTAAPTAAPTAAPTAAPTA--APSTVVPPATPPATAAPVPPTT---AIPT-- 533

Qy    439 EKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNT--PKEPAPTTPKEPAPTTPKE 496
            ||||     ||||    :|||   | ||    ||| |  |: |   |    |||
Db    534 --PAPTAAPTAAPTTAAPESPTTVTVP-PTAAPTAAPTTAVPEIPITVT---SAPTAAPT 587

Qy    497 PAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPK 556
           |||     ||||     |||:  |    |   ||   || ||                | |
Db    588 AAPTAAPTAAPTTAVPEIPTTVTSPPTAAPTTAAP---APNTT-----------VTVPP 632

Qy    557 GTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTS 616
           ||||   ||| |     ||     || |:                 || ||  | :|||
Db    633 TAAPTT-AAPAPNTTVTVPPTAAPTAAPPTVAH-------------APNTTAAPVTTTS 677

Qy    617 DKPAPTTPKETAPTTPKEPAPTTPKKP 643
           || |  |     |  | |  |  |
Db    678 ---APATTPEDDDIDP--PLPNDPINP 699


RESULT   4
US-08-928-361B-5
; Sequence 5, Application US/08928361B
; Patent No. 6071518
;  GENERAL INFORMATION:
;    APPLICANT:  Petersen, Carolyn
;    TITLE OF INVENTION:  PEPTIDES, POLYPEPTIDES, GLYCOPROTEINS,
;    TITLE OF INVENTION:  THEIR FUNCTIONAL MUTANTS, VARIANTS, ANALOGS AND FRAGMENTS
;    TITLE OF INVENTION:  FOR TREATMENT AND DETECTION/DIAGNOSIS OF CRYPTOSPORIDIUM
;    TITLE OF INVENTION:  SPECIES INFECTIONS
;    NUMBER OF SEQUENCES:  30
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  PETERS, VERNY, JONES & BIKSA
;      STREET:  385 Sherman Avenue, Suite 6
;      CITY:  Palo Alto
;      STATE:  CA
;      COUNTRY:  USA
;      ZIP:  94306-1840
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/08/928,361B
;      FILING DATE:  12-SEP-1997
;      CLASSIFICATION:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 60/026,062
;      FILING DATE:  13-SEP-1996
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Verny, Hana
;      REGISTRATION NUMBER:  30,518
;      REFERENCE/DOCKET NUMBER:  480.76-1(HV)
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  650-324-1677
;      TELEFAX:  650-324-1678
;  INFORMATION FOR SEQ ID NO:  5:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  1837 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
US-08-928-361B-5


  Query Match             9.8%;  Score 493;  DB 3;  Length 1837;
  Best Local Similarity   21.8%;  Pred. No. 1.4e-23;
```

Matches 226; Conservative 33; Mismatches 211; Indels 342; Gaps 40;

```
Qy    125 TSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAP-----TTP 179
Db      9 TALGLVAARPEVSDAEKNPALHEPHPDXP--PAEQXXLLPXEYDCTKFYYCEYGLKFIAP 66

Qy    180 KEPAPTT-----------TKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTK 227
Db     67 RDCAPGTEFKFSAQTCVHAALAGCTLPGPPAETT---QAPATTQ--APTTTQ--APTTTT 119

Qy    228 SAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTA 287
Db    120 QAPTTTTQATTTT----------QAPTTTQAPTTT----QATTTQAPTTTT-------- 156

Qy    288 PKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAP 347
Db    157 ---QAPTTTTQ-APTTTTQ-APTTTQ----------APTTTQ-APTTTQ--APTTITQAA 198

Qy    348 TTPKEPSPTTTKEPAPTTP--KEPAPTTPKKPAPTTPKEPAPTTPK-------------- 391
Db    199 TTPAATTPAAT-TPAATTPAATTPAATTPGVPAPTS----APVWPPICELLPNGCPADFD 253

Qy    392 ----------------------------------------------------EPAPTTTK 399
Db    254 IHLLIPHDKYCNLFYQCSNGYTFEQRCPEGLYFNPYVQRCDSPANVECDGEISPAPPVTE 313

Qy    400 K----------------PA---------------------------------------- 402
Db    314 GNEDEDIDIGDLLDNGCPANFEIDWLLPHGNRCDKYYQCVHGNLVERRCGAGTHFSFELQ 373

Qy    403 --------------------------------PTAP--KEPAPT----------TPKET 417
Db    374 QCDHIELVGCTLPGGESEEVDVDEDACTGWYCPTEPIEWEPLPNGCPADFSIDHLLPHES 433

Qy    418 ----------------------APTTPKKLTPTT----------PEKLAPTTPEKPAP 443
Db    434 DCGQYLQCVHGQTIARPCPGNLHFSPATQSCESPVTAGCQVFECDSDNQCTSTAAPTAAP 493

Qy    444 TTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPK 503
Db    494 TAAPTAAPTAAPTAAPSTVVPPA-TPPATAAPVPPTTAIPT----PAPTAAPTAAPTTAA 548

Qy    504 ETAPTT---PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAP 560
Db    549 PESPTTVTVPPTAAPTA----APTT----AVPEIPITVTSAPTAAPT--AAPTAAPTAAP 598

Qy    561 TTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTST-TSDKP 619
Db    599 TTAVPEIPTT-----VTSPPTAAPTT-AAPAPNT----------TVTVPPTAAPTTAAP 641

Qy    620 APTTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTP 679
Db    642 AP----NTTVTAPPTAAPTT-AAPAPNTTVTVPPTAAPTAAP-----PTVAH------AP 685

Qy    680 ELSAEPTPKALENSPKEPGVPTTKTPAATKPE 711
Db    686 NTTAAP------------VTTTSAPATT-PE 703
```

RESULT 2
5202236-25
;Patent No. 5202236
; APPLICANT: MAUGH, KATHY J.;ANDERSON, DAVID M.;STRAUSBERG,
;SUSAN L.;MCCANDLISS, RUSS;WEI, TENA;FILPULA, DAVID
; TITLE OF INVENTION: METHOD OF PRODUCING BIOADHESIVE
;PROTEIN
; NUMBER OF SEQUENCES: 39
; CURRENT APPLICATION DATA:
; APPLICATION NUMBER: US/07/528,762
; FILING DATE: 25-MAY-1990
; APPLICATION NUMBER: 82,456
; FILING DATE: 07-AUG-1987
; APPLICATION NUMBER: 933,945
; FILING DATE: 24-NOV-1986
; APPLICATION NUMBER: 650,128
; FILING DATE: 13-SEP-1984
;SEQ ID NO:25:
; LENGTH: 744
5202236-25

Query Match 10.1%; Score 506.5; DB 6; Length 744;
Best Local Similarity 29.9%; Pred. No. 7.8e-25;
Matches 239; Conservative 78; Mismatches 308; Indels 175; Gaps 50;

```
Qy     51 PKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKE 110
Db     24 PKMTYPPTYKPKPSYPP----------------TYKSKPTY----------KPKIT---- 53

Qy    111 TQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKE----PASTTPKE 166
Db     54 ----------YPPTYK--AKPS--------YPPTYKPKKTYPPTYKPKLTYPPTYKPKP 92

Qy    167 PTPTTIKSAPT-TPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPA--P 223
Db     93 SYPPTYKSKPTYKPKITYPPTYKAKPSYPPTYKPKKTYPPT-YKPKLTYPPTYKPKASYP 151

Qy    224 TTTKSAPTTPKEPAPTTPKKPAPT-TPKEPAPTT----PKEPTPTTPKEPAPTTKEP--- 275
Db    152 PTYKPKPSYP--PSYKTKKTYPPTYKPKLTYPPTYKPKPSYPPSYKPKKTYPPTYKPKLT 209

Qy    276 -APTTPKEPA--PTAPKKPA--PTTPKEPA--PTTPKEPA--PTTTKEPS--PTTPKEPA 324
Db    210 YPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPS 269

Qy    325 -PTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPA--PTTPKEPA----PTTPKKP 377
Db    270 YPSTYKAKPSYT---YPSTYKAKPTYP----PTYKAKPSYPPTYKAKPTYKAKPTYKAKP 322

Qy    378 A--PTTPKEPA--PTTPKEPA--PTTTKKPA--PTAPKEPAPTTPKETAPTTPKKLTPTT 429
Db    323 SYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKVKPTYKAKPTYPS 382

Qy    430 PEKLAPTTPE--KPAPTTPEELAPTTPEEPT--PTTPEEPA-PTTPKA--AAPNT----P 478
Db    383 TYKAKPSYPPTYKAKPSYP----PTYKAKPSYPPTYKAKPSYPPTYKAKPTYPSTYKAKP 438

Qy    479 KEPAPTTPKEPAPTT----PKEPAPTTPKETAPTT--PKGTAPTTLKEPAPTTPKKPAPK 532
Db    439 SYPPTYKPKISYPPTYKAKPSYPSTYKAKSSYPPTYKAKPSYPSTYKAKSSYPPTYKAKP 498

Qy    533 ELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPA- 591
Db    499 SYPPTYKAKPTYPSTYKAKPTYK--ARPSYP----PTYKAKPS------YPPTYKAKPSY 546

Qy    592 -PTTPKKPAPKELAPTTTKGPT--STTSDKPAPTTPKETAPTTPKEPAPTTPKKPA--PT 646
Db    547 PPTYKAKPTYK-AKPTYKAKPTYPSTYKAKPSYSPTYKAKPSYP----PTYKAKPSYPPT 601

Qy    647 ---TPETPPPTTSEVSTPTTTKE----PTTIHKSPD-----ESTPELSAEPT-PKALENS 693
Db    602 YKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTYPSTYKAK 661

Qy    694 PKEPGVPTTKTPAA------TKPEMTTTAKDKTTERDLRTTPETTTAAP--------KMT 739
Db    662 PSYP--PTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTYPSTYKAKPSYPPTHKAKPT 719

Qy    740 KETATTTEKTTESKITATTT 759
Db    720 YKAKPTYPSTYKAKPTYPST 739
```

RESULT 3
US-09-103-429A-3
; Sequence 3, Application US/09103429A
; Patent No. 6187558

GenCore version 4.5
Copyright (c) 1993 - 2000 Compugen Ltd.

OM protein - protein search, using sw model

Run on: April 26, 2002, 16:13:29 ; Search time 49.78 Seconds
(without alignments)
425.384 Million cell updates/sec

Title: US-09-556-246-1_COPY_200_1140
Perfect score: 5011
Sequence: 1 VKDNKKNRTKKKPTPKPPVV.....DMDYLPRVPNQGIIINPMLS 941

Scoring table: BLOSUM62
Gapop 10.0 , Gapext 0.5

Searched: 212252 seqs, 22503292 residues

Total number of hits satisfying chosen parameters: 212252

Minimum DB seq length: 0
Maximum DB seq length: 2000000000

Post-processing: Minimum Match 0%
Maximum Match 100%
Listing first 45 summaries

Database : Issued_Patents_AA:*
1: /cgn2_6/ptodata/2/iaa/5A_COMB.pep:*
2: /cgn2_6/ptodata/2/iaa/5B_COMB.pep:*
3: /cgn2_6/ptodata/2/iaa/6A_COMB.pep:*
4: /cgn2_6/ptodata/2/iaa/6B_COMB.pep:*
5: /cgn2_6/ptodata/2/iaa/PCTUS_COMB.pep:*
6: /cgn2_6/ptodata/2/iaa/backfiles1.pep:*

Pred. No. is the number of results predicted by chance to have a score greater than or equal to the score of the result being printed, and is derived by analysis of the total score distribution.

SUMMARIES

Result No.	Score	% Query Match	Length	DB	ID	Description
1	522	10.4	805	4	US-09-103-429A-4	Sequence 4, Appli
2	506.5	10.1	744	6	5202236-25	Patent No. 5202236
3	498.5	9.9	786	4	US-09-103-429A-3	Sequence 3, Appli
4	493	9.8	1837	3	US-08-928-361B-5	Sequence 5, Appli
5	488.5	9.7	826	1	US-07-638-431-2	Sequence 2, Appli
6	488.5	9.7	826	5	PCT-US92-00018-2	Sequence 2, Appli
7	476.5	9.5	652	6	5202236-13	Patent No. 5202236
8	466.5	9.3	1721	3	US-08-700-651-5	Sequence 5, Appli
9	463.5	9.2	1721	3	US-08-928-361B-6	Sequence 6, Appli
10	452	9.0	960	4	US-09-219-849-5	Sequence 5, Appli
11	423.5	8.5	1867	2	US-08-479-537A-5	Sequence 5, Appli
12	423.5	8.5	1867	4	US-09-083-116-5	Sequence 5, Appli
13	423.5	8.5	2035	2	US-08-479-537A-2	Sequence 2, Appli
14	423.5	8.5	2035	4	US-09-083-116-2	Sequence 2, Appli
15	419.5	8.4	2476	2	US-08-276-967-2	Sequence 2, Appli
16	419	8.4	1185	4	US-09-041-886-23	Sequence 23, Appl
17	417	8.3	829	1	US-08-642-255-132	Sequence 132, App
18	417	8.3	829	1	US-08-397-633A-53	Sequence 53, Appl
19	417	8.3	837	1	US-08-175-155-68	Sequence 68, Appl
20	417	8.3	837	1	US-08-477-509B-103	Sequence 103, App
21	417	8.3	837	1	US-08-642-255-101	Sequence 101, App
22	417	8.3	837	2	US-08-707-237A-75	Sequence 75, Appl
23	417	8.3	837	3	US-08-482-085B-103	Sequence 103, App
24	417	8.3	897	1	US-08-397-633A-50	Sequence 50, Appl
25	413.5	8.3	907	3	US-08-783-774-2	Sequence 2, Appli
26	413.5	8.3	907	5	PCT-US95-04611A-19	Sequence 19, Appl
27	404.5	8.1	408	1	US-07-609-716-65	Sequence 65, Appl
28	404.5	8.1	408	4	US-08-475-411A-65	Sequence 65, Appl
29	404.5	8.1	408	4	US-08-478-029A-65	Sequence 65, Appl
30	398.5	8.0	682	1	US-08-642-255-126	Sequence 126, App
31	398.5	8.0	682	1	US-08-397-633A-36	Sequence 36, Appl
32	381	7.6	1848	4	US-08-296-791-6	Sequence 6, Appli
33	381	7.6	1848	5	PCT-US95-10661A-6	Sequence 6, Appli
34	376	7.5	960	4	US-09-219-849-6	Sequence 6, Appli
35	374.5	7.5	1537	1	US-08-325-267A-2	Sequence 2, Appli
36	371	7.4	761	2	US-08-707-237A-84	Sequence 84, Appl
37	371	7.4	762	1	US-08-642-255-114	Sequence 114, App
38	371	7.4	762	1	US-08-397-633A-26	Sequence 26, Appl
39	370.5	7.4	1064	1	US-08-642-255-62	Sequence 62, Appl
40	369.5	7.4	1231	3	US-08-904-263A-4	Sequence 4, Appli
41	368	7.3	762	1	US-08-642-255-120	Sequence 120, App
42	368	7.3	762	1	US-08-397-633A-31	Sequence 31, Appl
43	367.5	7.3	1187	1	US-08-320-559-28	Sequence 28, Appl
44	367.5	7.3	1187	3	US-08-545-860D-28	Sequence 28, Appl
45	367.5	7.3	1187	5	PCT-US94-04496-28	Sequence 28, Appl

ALIGNMENTS

```
RESULT 1
US-09-103-429A-4
; Sequence 4, Application US/09103429A
; Patent No. 6187558
; GENERAL INFORMATION:
;   APPLICANT: Granados, Robert R
;   APPLICANT: Wang, Ping
;   TITLE OF INVENTION: A No. 6187558el Invertebrate Intestinal Mucin
;   TITLE OF INVENTION: cDNA and Related Products and Methods
;   NUMBER OF SEQUENCES: 4
;   CORRESPONDENCE ADDRESS:
;     ADDRESSEE: Brown, Pinnisi & Michaels, P.C.
;     STREET: 118 No. 6187558th Tioga
;     CITY: Ithaca
;     STATE: NY
;     COUNTRY: USA
;     ZIP: 14850
;   COMPUTER READABLE FORM:
;     MEDIUM TYPE: Floppy disk
;     COMPUTER: IBM PC compatible
;     OPERATING SYSTEM: PC-DOS/MS-DOS
;     SOFTWARE: PatentIn Release #1.0, Version #1.30
;   CURRENT APPLICATION DATA:
;     APPLICATION NUMBER: US/09/103,429A
;     FILING DATE: 24-JUN-1998
;     CLASSIFICATION: 800
;   ATTORNEY/AGENT INFORMATION:
;     NAME: Michaels, Christopher A
;     REGISTRATION NUMBER: 34,390
;     REFERENCE/DOCKET NUMBER: BTI-39
;   TELECOMMUNICATION INFORMATION:
;     TELEPHONE: (607) 256-2000
;     TELEFAX: (607) 256-3628
; INFORMATION FOR SEQ ID NO: 4:
;   SEQUENCE CHARACTERISTICS:
;     LENGTH: 805 amino acids
;     TYPE: amino acid
;     STRANDEDNESS: single
;     TOPOLOGY: linear
;   MOLECULE TYPE: protein
;   HYPOTHETICAL: NO
;   ANTI-SENSE: NO
;   ORIGINAL SOURCE:
;     ORGANISM: Trichoplusia ni
;     TISSUE TYPE: peritrophic membrane
US-09-103-429A-4

  Query Match          10.4%;  Score 522;  DB 4;  Length 805;
  Best Local Similarity  27.8%;  Pred. No. 8.9e-26;
```

```
          |. |    ||     |::  | :|      :|     |  |           |||:
Db    701 KKEEA----KEKKAAAPEEETPAKLGVKEEAKPKEKAEDAKA----------KEPSKP 744

Qy    570 TPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAP 629
          : ||     ||       ||  |  |:|   ||           ||  |     ||  |
Db    745 SEKE----KPK-------KEEVPAAPEKKDTKE---------EKTTESKKREEKPKMEAK 784

Qy    630 TTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDE 676
            ::       |  |  |  |    :| |: |  |:     |:|::
Db    785 AKEED-------KGLPQEPSKPKTEKAEKSSSTDQKDSQPSEKAPED 824
```

Search completed: April 26, 2002, 16:19:03
Job time: 328 sec

```
RESULT  15
NFH_RAT
ID   NFH_RAT        STANDARD;      PRT;   831 AA.
AC   P16884; Q63368;
DT   01-AUG-1990 (Rel. 15, Created)
DT   01-FEB-1994 (Rel. 28, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   NEUROFILAMENT TRIPLET H PROTEIN (200 KDA NEUROFILAMENT PROTEIN)
DE   (NEUROFILAMENT HEAVY POLYPEPTIDE) (NF-H) (FRAGMENT).
GN   NEFH OR NFH.
OS   Rattus norvegicus (Rat).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Rattus.
OX   NCBI_TaxID=10116;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Brain;
RX   MEDLINE=89065087; PubMed=3143606;
RA   Breen K.C., Robinson P.A., Wion D., Anderton B.H.;
RT   "Partial sequence of the rat heavy neurofilament polypeptide (NF-H).
RT   Identification of putative phosphorylation sites.";
RL   FEBS Lett. 241:213-218(1988).
RN   [2]
RP   SEQUENCE OF 37-831 FROM N.A.
RX   MEDLINE=88309090; PubMed=2457365;
RA   Dautigny A., Pham-Dinh D., Roussel C., Felix J.M., Nussbaum J.L.,
RA   Jolles P.;
RT   "The large neurofilament subunit (NF-H) of the rat: cDNA cloning and
RT   in situ detection.";
RL   Biochem. Biophys. Res. Commun. 154:1099-1106(1988).
RN   [3]
RP   SEQUENCE OF 1-89 AND 243-313 FROM N.A.
RX   MEDLINE=87080760; PubMed=2878828;
RA   Robinson P.A., Wion D., Anderton B.H.;
RT   "Isolation of a cDNA for the rat heavy neurofilament polypeptide
RT   (NF-H).";
RL   FEBS Lett. 209:203-205(1986).
RN   [4]
RP   SEQUENCE OF 318-831 FROM N.A.
RX   MEDLINE=89184647; PubMed=2928342;
RA   Lieberburg I., Spinner N., Snyder S., Anderson J., Goldgaber D.,
RA   Smulowitz M., Carroll Z., Emanuel B.S., Breitner J., Rubin L.;
RT   "Cloning of a cDNA encoding the rat high molecular weight
RT   neurofilament peptide (NF-H): developmental and tissue expression in
RT   the rat, and mapping of its human homologue to chromosomes 1 and
RT   22.";
RL   Proc. Natl. Acad. Sci. U.S.A. 86:2463-2467(1989).
CC   -!- FUNCTION: NEUROFILAMENTS USUALLY CONTAIN THREE IF PROTEINS: L, M,
CC       AND H WHICH ARE INVOLVED IN THE MAINTENANCE OF NEURONAL CALIBER.
CC       NF-H HAS AN IMPORTANT FUNCTION IN MATURE AXONS THAT IS NOT
CC       SUBSERVED BY THE TWO SMALLER NF PROTEINS.
CC   -!- PTM: THERE ARE A NUMBER OF REPEATS OF THE TRIPEPTIDE K-S-P, NFH IS
CC       PHOSPHORYLATED ON A NUMBER OF THE SERINES IN THIS MOTIF. IT IS
CC       THOUGHT THAT PHOSPHORYLATION OF NFH RESULTS IN THE FORMATION OF
CC       INTERFILAMENT CROSS BRIDGES THAT ARE IMPORTANT IN THE MAINTENANCE
CC       OF AXONAL CALIBER.
CC   -!- PTM: PHOSPHORYLATION SEEMS TO PLAY A MAJOR ROLE IN THE FUNCTIONING
CC       OF THE LARGER NEUROFILAMENT POLYPEPTIDES (NF-M AND NF-H), THE
CC       LEVELS OF PHOSPHORYLATION BEING ALTERED DEVELOPMENTALLY AND
CC       COINCIDENT WITH A CHANGE IN THE NEUROFILAMENT FUNCTION.
CC   -!- SIMILARITY: BELONGS TO THE INTERMEDIATE FILAMENT FAMILY.
CC   -!- CAUTION: REF.1 SEQUENCE DIFFERS FROM THAT SHOWN FROM POSITION 783
CC       ONWARD AND IS LONGER DUE TO A FRAMESHIFT.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M37227; AAA41693.1; ALT_FRAME.
DR   EMBL; X13804; CAA32038.1; ALT_FRAME.
DR   EMBL; M21964; AAA41695.1; -.
DR   EMBL; J04517; AAA41692.1; -.
DR   PIR; A30796; A30796.
DR   PIR; A25649; A25649.
DR   PIR; B25649; B25649.
DR   PIR; S02003; S02003.
DR   InterPro; IPR001664; IF.
DR   Pfam; PF00038; filament; 1.
DR   PROSITE; PS00226; IF; 1.
KW   Intermediate filament; Coiled coil; Neurone; Phosphorylation; Repeat.
FT   NON_TER       1      1
FT   DOMAIN      276    641       51 X 3 AA TANDEM REPEATS OF K-S-P.
FT   CONFLICT    164    164       L -> I (IN REF. 2).
FT   CONFLICT    185    185       I -> S (IN REF. 2).
FT   CONFLICT    193    193       L -> T (IN REF. 2).
FT   CONFLICT    199    199       M -> T (IN REF. 2).
FT   CONFLICT    346    346       K -> N (IN REF. 1).
FT   CONFLICT    373    373       A -> V (IN REF. 1 AND 4).
FT   CONFLICT    482    482       G -> E (IN REF. 2 AND 4).
FT   CONFLICT    485    485       P -> S (IN REF. 2).
FT   CONFLICT    570    571       RK -> KE (IN REF. 2 AND 4).
FT   CONFLICT    591    591       P -> T (IN REF. 2 AND 4).
FT   CONFLICT    727    727       A -> V (IN REF. 4).
FT   CONFLICT    757    759       AAP -> GST (IN REF. 4).
FT   CONFLICT    769    769       T -> L (IN REF. 2).
FT   CONFLICT    775    775       R -> P (IN REF. 2 AND 4).
SQ   SEQUENCE   831 AA;  89486 MW;  1B0973C3F13EF768 CRC64;

  Query Match             9.9%;  Score 497;  DB 1;  Length 831;
  Best Local Similarity   27.6%;  Pred. No. 2.3e-13;
  Matches  195;  Conservative   59;  Mismatches  307;  Indels  146;  Gaps   35;

Qy      37 PDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQ 96
           |  ||  | || :  ||   :             : ||| :   :   ::  |  |  :
Db     197 PSMST--HIKVKSEEKIKVVE-----------KSEKETVIVEEQTEEIQVTEEVTEEED 242

Qy      97 TSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEP----T 152
             |:|:  : :  :    ||     | ::  | |    |  || |     | |    :
Db     243 KEAQGEEEEEAEEGGEEAATTS-----PPAEEAASP----EKETKSPVKEEAKSPAEAKS 293

Qy     153 PTTPKEPA-STTPKEPTPTTIKSAPTTPKEPA----PTTTKSAPTTPKEPA----PTTTK 203
           |   | || : :| |     :  :|   | ||     |   || |   | ||    | | |
Db     294 PAEAKSPAEAKSPAEVKSPAVAKSPAEVKSPAEVKSPAEAKS-PAEAKSPAEVKSPATVK 352

Qy     204 EP----APTTPKEPA----PTTTKEPA----PTTTKSAPTTPKEPA----PTTPKKPA-- 245
           |    :|   | ||     |   | ||     | : || |   | ||    |   | ||
Db     353 SPGEAKSPAEAKSPAEVKSPAEAKSPAEAKSPASVKS-PGEAKSPAEAKSPAEVKSPATV 411

Qy     246 --PTTPKEPA----PTTPKEP----TPTTPKEPAPTTKEP----APTTPKEPA----PTA 287
             |   | ||    | | | |    :|   | || : | |    :|   | ||    |
Db     412 KSPVEAKSPAEVKSPVTVKSPAEAKSPVEVKSPA-SVKSPSEAKSPAGAKSPAEAKSPVV 470

Qy     288 PKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPA----PTTTKS-----APTTTKEP 338
           | ||    | ||   |   | : :  ||   | ||    |   ||     :|   | |
Db     471 AKSPAEA--KSPAGAKPPAEAKSPAEAKSPAEAKSPAEAKSPAEAKSPVEVKSPEKAKSP 528

Qy     339 APTTTKS-----APTTPKEPSPTTTKEPA----PTTPKEPAPTTPKKPAPTTPKEPAPTT 389
           |      :|   | |    | ||    |   | |     | |     :
Db     529 VKEGAKSLAEAKSPEKAKSPVKEEIKPPAEVKSPEKAKSPMRKEAKSPEKAKTLDVKSPE 588

Qy     390 PKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEEL 449
           | ||    |:||     |   :  || | :  |: | |  ||:||   |  :|   ||:
Db     589 AKPPAKEEAKRPADIRSPEQVKSPAKEEAKSPEKEETRT--EKVAPKKEEVKSPV--EEV 644

Qy     450 APTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTT 509
           :||  `  ||  | |||     : |: ||   ::|     |||    ||::
Db     645 ---KAKEPPKKVEEEKTPATPKTEVKESKKDEAPKEAQKP-KAEEKEPLTEKPKDSPGEA 700

Qy     510 PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPT 569
```

```
          :|||:   :|||:    :|||:   :|||:|   |||:    :|||:|   :|||:|   :|||:|
Db   1672 -SPTSPSYSPTSPSY-SPTSPSYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY-SPTSP 1725

Qy    383 KEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPA 442
           :||:|   :|||:    :|||:|   :||:|  | ||:|   :|||:|   :|||:|
Db   1726 SY-SPTSPSY-SPTSPSY-SPTSPNY-SPTSPNYT-PTSP-SYSPTSP-SYSPTSPNY-T 1777

Qy    443 PTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEP--APT 500
           ||:|   :|||:|   :||:|   :|||:| : :|::|:     ||: |   ||  |  :|:
Db   1778 PTSP-NYSPTSPSY-SPTSPSY-SPTSP-SYSPSSPR----YTPQSPT-YTPSSPSYSPS 1828

Qy    501 TPKETAPTTPKGTAPTTLKEPAPTTPK-KPAPKELAPTTTKEPTSTTSDKPAPTTPKGTA 559
           :|   :||:|| |          ||:|   |:  |  ||: |   | || | :||:||  :
Db   1829 SP-SYSPTSPKYT---------PTSPSYSPSSPEYTPTSPK--YSPTSPKYSPTSPK-YS 1875

Qy    560 PTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPT-STTSDK 618
           ||:|   :|||||  :||:|  |   |     ||:||       :||:   || | || |
Db   1876 PTSPTY-SPTTPKY-SPTSP--TYSPTSPVYTPTSPK------YSPTS---PTYSPTSPK 1922

Qy    619 PAPTTPKETAPTTPKEP--APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDE 676
           :||:|   :||:||    :||:|   :||:|      |  :::|           |||:
Db   1923 YSPTSP-TYSPTSPKGSTYSPTSPGY-SPTSP------TYSLTSPAI---------SPDD 1965

Qy    677 STPE 680
           |  |
Db   1966 SDEE 1969
```

```
RESULT  14
RPB1_MOUSE
ID   RPB1_MOUSE     STANDARD;      PRT;  1970 AA.
AC   P08775;
DT   01-NOV-1988 (Rel. 09, Created)
DT   01-MAR-1992 (Rel. 21, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   DNA-DIRECTED RNA POLYMERASE II LARGEST SUBUNIT (EC 2.7.7.6) (RPB1).
GN   POLR2A OR RPO2-1 OR RPII215.
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=87280135; PubMed=3038894;
RA   Ahearn J.M. Jr., Bartolomei M.S., West M.L., Cisek L.J., Corden J.L.;
RT   "Cloning and sequence analysis of the mouse genomic locus encoding
RT   the largest subunit of RNA polymerase II.";
RL   J. Biol. Chem. 262:10695-10705(1987).
RN   [2]
RP   SEQUENCE OF 1587-1970 FROM N.A.
RX   MEDLINE=86068017; PubMed=2999785;
RA   Corden J.L., Cadena D.L., Ahearn J.M. Jr., Dahmus M.E.;
RT   "A unique structure at the carboxyl terminus of the largest subunit
RT   of eukaryotic RNA polymerase II.";
RL   Proc. Natl. Acad. Sci. U.S.A. 82:7934-7938(1985).
RN   [3]
RP   REVISIONS, AND PRESENCE OF AN ADDITIONAL EXON.
RX   MEDLINE=92178992; PubMed=1542581;
RA   Wintzerith M., Acker J., Vicaire S., Vigneron M., Kedinger C.;
RT   "Complete sequence of the human RNA polymerase II largest subunit.";
RL   Nucleic Acids Res. 20:910-910(1992).
CC   -!- FUNCTION: DNA-DEPENDENT RNA POLYMERASE CATALYZES THE TRANSCRIPTION
CC       OF DNA INTO RNA USING THE FOUR RIBONUCLEOSIDE TRIPHOSPHATES AS
CC       SUBSTRATES.
CC   -!- CATALYTIC ACTIVITY: N NUCLEOSIDE TRIPHOSPHATE = N PYROPHOSPHATE +
CC       RNA(N).
CC   -!- SUBUNIT: RNA POLYMERASE II CONSISTS OF 12 DIFFERENT SUBUNITS.
CC   -!- SUBCELLULAR LOCATION: NUCLEAR.
CC   -!- PTM: THE TANDEM 7 RESIDUES REPEATS CAN BE HIGHLY PHOSPHORYLATED.
CC       THE PHOSPHORYLATION ACTIVATES POL2.
CC   -!- MISCELLANEOUS: THREE DISTINCT ZINC-CONTAINING RNA POLYMERASES ARE
CC       FOUND IN EUKARYOTIC NUCLEI: POLYMERASE I FOR THE RIBOSOMAL RNA
CC       PRECURSOR, POLYMERASE II FOR THE MRNA PRECURSOR, AND POLYMERASE
CC       III FOR 5S AND TRNA GENES.
CC   -!- SIMILARITY: BELONGS TO THE RNA POLYMERASE BETA' CHAIN FAMILY.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M12130; AAA40071.1; -.
DR   EMBL; M14101; AAA40071.1; JOINED.
DR   PIR; A28490; A28490.
DR   MGD; MGI:98086; Rpo2-1.
DR   InterPro; IPR000684; RNA_polII_repeat.
DR   InterPro; IPR000722; RNA_pol_A.
DR   InterPro; IPR002879; RNA_pol_A2.
DR   Pfam; PF00623; RNA_pol_A; 1.
DR   Pfam; PF01854; RNA_pol_A2; 1.
DR   PROSITE; PS00115; RNA_POL_II_REPEAT; 42.
KW   Transferase; DNA-directed RNA polymerase; Transcription; Zinc; Repeat;
KW   DNA-binding; Nuclear protein; Phosphorylation; Zinc-finger.
FT   ZN_FING      71     87       C2H2-TYPE (POTENTIAL).
FT   DOMAIN     1590   1958       CARBOXYL-TERMINAL 7-RESIDUE REPEATS.
FT   CONFLICT   1498   1498       P -> R (IN REF. 1 AND 2).
FT   CONFLICT   1499   1536       MISSING (IN REF. 1 AND 2).
SQ   SEQUENCE   1970 AA;  217175 MW;  7D76F38FD92A657E CRC64;
```

```
  Query Match             9.9%;  Score 498.5;  DB 1;  Length 1970;
  Best Local Similarity   33.1%;  Pred. No. 3.9e-13;
  Matches  180;  Conservative  101;  Mismatches  168;  Indels   95;  Gaps   47;

Qy    151 PTPTTPKEPASTTPKE-PTPTTIKSAPTTPKEPAPTTTKSAPTTPKE-----PAPTTTKE 204
          |:|     || |   :  ||     :|:      |     :|:   :     |  :
Db   1507 PSPMGGISPAMTPWNQGATPAYGAWSPSVGSGMTPGAAGFSPSAASDASGFSPGYSPAWS 1566

Qy    205 PAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEP--APTTPKEPTP 262
          | | :|  | |::   |:|     |   :|  ||   |: |   ||: |  :||:|   :|
Db   1567 PTPGSPGSPGPSSSPYIPSPGGAMSPSYSPTSPA-YEPRSPGGYTPQSPSYSPTSPSY-SP 1624

Qy    263 TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKE 322
          |:|   :|||:   :|||:|   :|||:|   :||:|   :||:|            |||:|
Db   1625 TSPSY-SPTSPNYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY---------SPTSPSY 1671

Qy    323 PAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTP 382
           :|||:   :|||:    :|||:   :|||:|   |||:    :|||:|   :|||:|   :|||:|
Db   1672 -SPTSPSYSPTSPSY-SPTSPSYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY-SPTSP 1725

Qy    383 KEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPA 442
           :|||:|   :|||:    :|||:|   :|||:|  | ||:|   :|||:|   :|||:|
Db   1726 SY-SPTSPSY-SPTSPSY-SPTSPNY-SPTSPNYT-PTSP-SYSPTSP-SYSPTSPNY-T 1777

Qy    443 PTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEP--APT 500
           ||:|   :|||:|   :|||:|   :|||:| : :|::|:     ||: |   ||  |  :|:
Db   1778 PTSP-NYSPTSPSY-SPTSPSY-SPTSP-SYSPSSPR----YTPQSPT-YTPSSPSYSPS 1828

Qy    501 TPKETAPTTPKGTAPTTLKEPAPTTPK-KPAPKELAPTTTKEPTSTTSDKPAPTTPKGTA 559
           :|   :||:|| |          ||:|   |:  |  | : |   | || | :||:||  :
Db   1829 SP-SYSPTSPKYT---------PTSPSYSPSSPEYTPASPK--YSPTSPKYSPTSPK-YS 1875

Qy    560 PTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPT-STTSDK 618
           ||:|   :|||||  :||:|  |   |     ||:||       :||:   || | || |
Db   1876 PTSPTY-SPTTPKY-SPTSP--TYSPTSPVYTPTSPK------YSPTS---PTYSPTSPK 1922

Qy    619 PAPTTPKETAPTTPKEP--APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDE 676
           :||:|   :||:||    :||:|   :||:|      |  :::|           |||:
Db   1923 YSPTSP-TYSPTSPKGSTYSPTSPGY-SPTSP------TYSLTSPAI---------SPDD 1965

Qy    677 STPE 680
           |  |
Db   1966 SDEE 1969
```

```
Qy      92 TTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEP 151
           ||    :  ||: | :|: : : |||:|       :| | ||  |      |    ||
Db       3 TTAAAVAATGKDTTAAAEGSAAAEKTAA-----AGEVSAPPT--AAVAATGEDATT---- 51

Qy     152 TPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPK 211
              |: |    |      ||||   || |    ||||    ||||    ||||
Db      52 --------AAATAAAETTAAAGEAPTTTTAPATTAAGKAPTTAAATAPTTAAAGAPTTAT 103

Qy     212 EPAPTTTKEPAPTTTKS-APTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPA- 269
           || |   | |||  | ||||      :|    ||||    | :  :  : :: :|
Db     104 GKAPATAAAPVPTTAASKAPTTAAAATHSTAAAAAPTTAASAAKSKERSTSSSSEEEHCH 163

Qy     270 --PTTKEPAPT---TPKEPAPTAPKKPAPTTPK----------EPAPTTPKEPAPTTTKE 314
             |: :|   :   | |:      ||     ||          :|   : :|   ||||
Db     164 VKPSKREMCGSKGITKKQ----CKKKNCCFDPKGHGGIHCFHRKPKGHSHEEHTTTTTK- 218

Qy     315 PSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTP 374
              |||| : | |||    ||||             ||| :  |||
Db     219 ---------APTTIQIATTTT---TPTTT-----------TTTTKATPTT--------- 245

Qy     375 KKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLA 434
                    ||  :  ||||                 || | |  ||    ||||
Db     246 ------------TTTTKATPTTT------------TTTKATTTTT----TPTT----- 270

Qy     435 PTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTP 494
                 ||    | |||   | |||
Db     271 ---------TTTTTKATTTPTTTTTTTP------------------------------- 289

Qy     495 KEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTT 554
                  |  ||   |  ||        | |           :|    :    :
Db     290 ----------TTTTTKATTTTTTTSGECKMEPSK----------REDCGYSGITESQCR 328

Qy     555 PKG-----TAPTT-----PKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELA 604
           ||     : | |              ||:     |    |    |    :|    : :
Db     329 TKGCCFDSSIPQTKWCFYTLSQVADCKVEPSQRVDCGFRGIT----ADQCRQKNCCFDSS 384

Qy     605 PTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTT 664
            : ||    :||   |  |   | |||   |  ||  |   |||       |:  :|||||
Db     385 ISGTKWCFYSTSQVAA--TKTTTTPTTTTTPTTTTTTKATTTTP------TTTTTTPTTT 436

Qy     665 KEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERD 724
              ||                 | ||          || ||  | |  |||   | |
Db     437 TTTTT---------------TTKA----------TTTTPTTTTPTTTTT---KAT--- 463

Qy     725 LRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAP 784
             ||| |||  |     | |||: |        |||  |:|||  ||  | ||   ||
Db     464 -TTTPTTTTTTP-----TTTTKAT-------TTTPTTTTTTPTTTTTKATTTTPTTTTT 510

Qy     785 KVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPK----T 840
             |||| | |||    | |    |  ||        : |  :|     :|  : |    :
Db     511 TTTTTKATTTTTSGECKME----PSKRADCGYPGITESQCRSKGCCFDSSIPQTKWCFYS 566

Qy     841 MPRVRKPKTTPTPR 854
           :|:|   |  |: |
Db     567 LPQVADCKVAPSSR 580
```

```
RESULT  13
RPB1_HUMAN
ID   RPB1_HUMAN     STANDARD;      PRT;  1970 AA.
AC   P24928;
DT   01-MAR-1992 (Rel. 21, Created)
DT   01-MAR-1992 (Rel. 21, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   DNA-DIRECTED RNA POLYMERASE II LARGEST SUBUNIT (EC 2.7.7.6) (RPB1).
GN   POLR2A.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=92178992; PubMed=1542581;
RA   Wintzerith M., Acker J., Vicaire S., Vigneron M., Kedinger C.;
RT   "Complete sequence of the human RNA polymerase II largest subunit.";
RL   Nucleic Acids Res. 20:910-910(1992).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=95347616; PubMed=7622068;
RA   Mita K., Tsuji H., Morimyo M., Takahashi E., Nenoi M.,
RA   Ichimura S., Yamauchi M., Hongo E., Hayashi A.;
RT   "The human gene encoding the largest subunit of RNA polymerase II.";
RL   Gene 159:285-286(1995).
CC   -!- FUNCTION: DNA-DEPENDENT RNA POLYMERASE CATALYZES THE TRANSCRIPTION
CC       OF DNA INTO RNA USING THE FOUR RIBONUCLEOSIDE TRIPHOSPHATES AS
CC       SUBSTRATES.
CC   -!- CATALYTIC ACTIVITY: N NUCLEOSIDE TRIPHOSPHATE = N PYROPHOSPHATE +
CC       RNA(N).
CC   -!- SUBUNIT: RNA POLYMERASE II CONSISTS OF 12 DIFFERENT SUBUNITS.
CC   -!- SUBCELLULAR LOCATION: NUCLEAR.
CC   -!- PTM: THE TANDEM 7 RESIDUES REPEATS CAN BE HIGHLY PHOSPHORYLATED.
CC       THE PHOSPHORYLATION ACTIVATES POL2.
CC   -!- MISCELLANEOUS: THREE DISTINCT ZINC-CONTAINING RNA POLYMERASES ARE
CC       FOUND IN EUKARYOTIC NUCLEI: POLYMERASE I FOR THE RIBOSOMAL RNA
CC       PRECURSOR, POLYMERASE II FOR THE MRNA PRECURSOR, AND POLYMERASE
CC       III FOR 5S AND TRNA GENES.
CC   -!- SIMILARITY: BELONGS TO THE RNA POLYMERASE BETA' CHAIN FAMILY.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X63564; CAA45125.1; -.
DR   EMBL; X74874; CAA52862.1; -.
DR   EMBL; X74873; CAA52862.1; JOINED.
DR   EMBL; X74872; CAA52862.1; JOINED.
DR   EMBL; X74871; CAA52862.1; JOINED.
DR   EMBL; X74870; CAA52862.1; JOINED.
DR   PIR; S21054; S21054.
DR   MIM; 180660; -.
DR   InterPro; IPR000684; RNA_polII_repeat.
DR   InterPro; IPR000722; RNA_pol_A.
DR   InterPro; IPR002879; RNA_pol_A2.
DR   Pfam; PF00623; RNA_pol_A; 1.
DR   Pfam; PF01854; RNA_pol_A2; 1.
DR   PROSITE; PS00115; RNA_POL_II_REPEAT; 43.
KW   Transferase; DNA-directed RNA polymerase; Transcription; Zinc; Repeat;
KW   DNA-binding; Nuclear protein; Phosphorylation; Zinc-finger.
FT   ZN_FING      71     87       C2H2-TYPE (POTENTIAL).
FT   DOMAIN     1590   1958       CARBOXYL-TERMINAL 7-RESIDUE REPEATS.
FT   CONFLICT   1067   1067       W -> L (IN REF. 2).
FT   CONFLICT   1449   1449       D -> Y (IN REF. 2).
SQ   SEQUENCE   1970 AA;  217205 MW;  6876FC25692A657E CRC64;
```

```
  Query Match           10.0%;  Score 503.5;  DB 1;  Length 1970;
  Best Local Similarity  33.3%;  Pred. No. 2.5e-13;
  Matches  181;  Conservative  101;  Mismatches  167;  Indels   95;  Gaps   47;

Qy     151 PTPTTPKEPASTTPKE-PTPTTIKSAPTTPKEPAPTTTKSAPTTPKE-----PAPTTTKE 204
           |:|      || |   :  ||     :|:       |     :|:   :     |  :
Db    1507 PSPMGGISPAMTPWNQGATPAYGAWSPSVGSGMTPGAAGFSPSAASDASGFSPGYSPAWS 1566

Qy     205 PAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEP--APTTPKEPTP 262
           | | :|  | |::   |:|    |   :|  ||   |: |   ||: |  :||:|   :|
Db    1567 PTPGSPGSPGPSSPYIPSPGGAMSPSYSPTSPA-YEPRSPGGYTPQSPSYSPTSPSY-SP 1624

Qy     263 TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKE 322
           |:|   :||:   :||:|   :||:|   :||:|   :||:|           |||:|
Db    1625 TSPSY-SPTSPNYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY---------SPTSPSY 1671

Qy     323 PAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTP 382
```

```
Db    470 PTYKPKPSYPPTYKPKITYPPTYKRKPSYPTPYKQKPSYPPIYKSKSSYPTSYKSKKTYP 529

Qy    508 TT--PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPT-TPKGTAPTTPK 564

Db    530 PTYKPKITYPPTYK------PKPSYPPSYKPKTTYPPTYKPKIRYPPTYKPKASYPPTYK 583

Qy    565 EPAPTTPKEPAPTTPKGTAPTTLKEPAPTTP----KKPAPKELAPTTTKGPTSTTSDKPA 620

Db    584 --------------PKITYPPTYK-PKPSYPTPYKQKPSYPPIYKSKSSYPTAYKSKKTY 628

Qy    621 PTT--PKETAPTTPKEPAPTTPK--KPAPTTPET-------PPPTTSEVSTPTTTKEPTT 669

Db    629 PPTYKPKITYPPTYK-PKPSYPPSYRPKITYPPTYKPKKSYPQAYKSKGSYPPSYQPKKT 687

Qy    670 IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKT----TERDL 725

Db    688 Y---PPSYKPKKTYPPTYK-----PKISYPPTYK----TKPSYPASYKRKTSYPPTYKPK 735

Qy    726 RTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPK 785

Db    736 ISYPSTYKAKP-----SYPPTYKPKPSYASSYKPKIRYPPTYKPKPSYASSYK-----PK 785

Qy    786 VTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ---KPTKAPKKP-TSTKKPK-T 840

Db    786 I--------------RYPPTYKPKP----SYASSYKPKIRYPPTYKPKPSYASSYKPKIT 827

Qy    841 MPRVRKPKTT--PTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQ 882

Db    828 YPPTYKPKISYPPTYKPKITYPPTYKPKISYPPAYKPKISYPSQ 871


RESULT  12
MUC1_XENLA
ID   MUC1_XENLA     STANDARD;      PRT;   662 AA.
AC   Q05049;
DT   01-OCT-1994 (Rel. 30, Created)
DT   01-OCT-1994 (Rel. 30, Last sequence update)
DT   01-OCT-1994 (Rel. 30, Last annotation update)
DE   INTEGUMENTARY MUCIN C.1 (FIM-C.1) (FRAGMENT).
OS   Xenopus laevis (African clawed frog).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Amphibia; Batrachia; Anura; Mesobatrachia; Pipoidea; Pipidae;
OC   Xenopodinae; Xenopus.
OX   NCBI_TaxID=8355;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Skin;
RX   MEDLINE=93077556; PubMed=1447205;
RA   Hauser F., Hoffmann W.;
RT   "P-domains as shuffled cysteine-rich modules in integumentary mucin
RT   C.1 (FIM-C.1) from Xenopus laevis. Polydispersity and genetic
RT   polymorphism.";
RL   J. Biol. Chem. 267:24620-24624(1992).
CC   -!- FUNCTION: COULD BE INVOLVED IN DEFENSE AGAINST MICROBIAL
CC       INFECTIONS. PROTECTS THE EPITHELIA FROM EXTERNAL ENVIRONMENT.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- ALTERNATIVE PRODUCTS: A NUMBER OF DIFFERENT FORMS OF THE PROTEIN
CC       MAY BE PRODUCED BY ALTERNATIVE SPLICING.
CC   -!- TISSUE SPECIFICITY: SKIN.
CC   -!- PTM: EXTENSIVELY O-GLYCOSYLATED.
CC   -!- SIMILARITY: CONTAINS 6 P-TYPE (TREFOIL) DOMAINS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; L02115; AAA74725.1; -.
DR   PIR; A45155; A45155.
DR   HSSP; P04002; 1WFA.
DR   InterPro; IPR000519; P_trefoil.
DR   Pfam; PF00088; trefoil; 6.
DR   SMART; SM00018; P; 6.
DR   PROSITE; PS00025; P_TREFOIL; 6.
KW   Repeat; Amphibian skin; Glycoprotein; Alternative splicing.
FT   NON_TER       1      1
FT   DOMAIN       81    144       8 X 8 AA APPROXIMATE TANDEM REPEATS,
FT                                ALA/THR-RICH.
FT   REPEAT       81     88       1-1.
FT   REPEAT       89     96       1-2.
FT   REPEAT       97    104       1-3.
FT   REPEAT      105    112       1-4.
FT   REPEAT      113    120       1-5.
FT   REPEAT      121    128       1-6.
FT   REPEAT      129    136       1-7.
FT   REPEAT      137    144       1-8.
FT   DOMAIN      161    202       P-TYPE 1.
FT   DOMAIN      218    301       8 X APPROXIMATE TANDEM REPEATS, THR-RICH.
FT   REPEAT      218    224       2-1.
FT   REPEAT      225    239       2-2.
FT   REPEAT      240    249       2-3.
FT   REPEAT      250    259       2-4.
FT   REPEAT      260    275       2-5.
FT   REPEAT      276    287       2-6.
FT   REPEAT      288    294       2-7.
FT   REPEAT      295    301       2-8.
FT   DOMAIN      306    347       P-TYPE 2.
FT   DOMAIN      353    394       P-TYPE 3.
FT   DOMAIN      402    522       12 X APPROXIMATE TANDEM REPEATS,
FT                                THR-RICH.
FT   REPEAT      402    411       3-1.
FT   REPEAT      412    419       3-2.
FT   REPEAT      420    431       3-3.
FT   REPEAT      432    443       3-4.
FT   REPEAT      444    453       3-5.
FT   REPEAT      454    460       3-6.
FT   REPEAT      461    472       3-7.
FT   REPEAT      473    479       3-8.
FT   REPEAT      480    491       3-9.
FT   REPEAT      492    498       3-10.
FT   REPEAT      499    515       3-11.
FT   REPEAT      516    522       3-12.
FT   DOMAIN      525    566       P-TYPE 4.
FT   DOMAIN      572    613       P-TYPE 5.
FT   DOMAIN      620    661       P-TYPE 6.
FT   DISULFID    162    188       BY SIMILARITY.
FT   DISULFID    172    187       BY SIMILARITY.
FT   DISULFID    182    199       BY SIMILARITY.
FT   DISULFID    307    333       BY SIMILARITY.
FT   DISULFID    317    332       BY SIMILARITY.
FT   DISULFID    327    344       BY SIMILARITY.
FT   DISULFID    354    380       BY SIMILARITY.
FT   DISULFID    364    379       BY SIMILARITY.
FT   DISULFID    374    391       BY SIMILARITY.
FT   DISULFID    526    552       BY SIMILARITY.
FT   DISULFID    536    551       BY SIMILARITY.
FT   DISULFID    546    563       BY SIMILARITY.
FT   DISULFID    573    599       BY SIMILARITY.
FT   DISULFID    583    598       BY SIMILARITY.
FT   DISULFID    593    610       BY SIMILARITY.
FT   DISULFID    621    647       BY SIMILARITY.
FT   DISULFID    631    646       BY SIMILARITY.
FT   DISULFID    641    658       BY SIMILARITY.
FT   VARIANT     276    276       K -> E.
FT   VARIANT     354    354       C -> R.
FT   VARIANT     415    415       T -> A.
SQ   SEQUENCE   662 AA;  67774 MW;  F085277F1ED2FD40 CRC64;


  Query Match             10.2%;  Score 509.5;  DB 1;  Length 662;
  Best Local Similarity   28.1%;  Pred. No. 6.3e-14;
  Matches  223;  Conservative   48;  Mismatches  276;  Indels  247;  Gaps   30;
```

```
Db    117 PVAATPTPVVQIPVAAPVIATPPVAASAPTPAAVTPVISPVIASPPVVPANTTVPVAAPV 176

Qy    280 PKEPA--PTAPKKPAPTTPKEPAPTTPKEPAPTTTKE-PSPTTPKEPAP-----TTTKSA 331

Db    177 AAVPAAVPVVAPVLAPAVAPAVAPVVAETPAPPPVAEIPVATIPECVAPLIPEVSVVATK 236

Qy    332 PTTTKEP---APTTTKSAPTTPKEPSPTTTKEPA-PTTPKEPAPTTPKKPAPTTPKEPAP 387

Db    237 PLAAAEPVVVAPPATETPVVAPAAASPHVSVAPAVETAVVAPVSASTEPPVAAATLTTAP 296

Qy    388 TTP--------KEPAPTTTKKPAPTAPKEP---APTTPKET--------APTTPKKLTPT 428

Db    297 ETPALAPVVAESQVAANTVVATPPTPAPEPETIAPPVVAETPEVASVAVAETTPPVVPPV 356

Qy    429 TPEKL-APTTPEKPAPTTPEELAPTTPEEPTPTTPEEP---APTTPKAAAPNTPKEPAPT 484

Db    357 AAESIPAPVVATTPVPAT---LAVTDPDVTASAVPELPPVIAPSPVPSAVAETPVDLAPP 413

Qy    485 TPKEPAPTTPKEPAPTTPKETAPTTP-KGTAPTTLKE----------PAPTTPKKPAPKE 533

Db.   414 V----LPPVAAEPVPAVVAEETPETPAPASAPVTIAALDIPEVAPVIAAPSDAPAEAPSA 469

Qy    534 LAPTTTKEPTSTTSDKPAPTTPKGTAPTTP------KEPAPTTPKEP--APTTPKGTAPT 585

Db    470 AAPIVSTPP--TTASVPETTAPPAAVPTEPIDVSVLSEAAIETPVAPPVEVTTEVAVADV 527

Qy    586 TLKEPA------PTTPKKPAPKELAPTTT----KGPTSTTSDKPAPTT----PKETAPTT 631

Db    528 APPEAAADLIIEPVEPPAPIPDLLEQTTSVPAVEAAESTSS--PIPETSLPPPNEAVASP 585

Qy    632 PKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPD----------ESTPEL 681

Db    586 EVAVAPITAPEPIP----EPEPSLATPTEPIPVEAPVVIQEAVDAVEVPVTETSTSIPET 641

Qy    682 SAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTP--ETTTAAPKMT 739

Db    642 TVE-FPEAVAEKVLDPAI--TEAPVTTQEPDVANINDGAPATEI-TTPAVEIVTAAAEVS 697

Qy    740 KETATTTEKTTESKITATTTQVTST------TTQDTTPFKITTLKTTTLA-PKVTTTKKT 792

Db    698 DIAIPVIDPPVPQEIAVAEIPETDTKPAEVIVEQSTIPIEAPVPEVSKYAEPVISEAPAA 757

Qy    793 ITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPT 852

Db    758 EVPITAGDNPDNTSVGISEVVPTIAEKPVEEVPTS--EIPEQSSSPS--DSVPVAKITPL 813

Qy    853 PRKMTSTMPELNPTSRIAEAM---LQTTTRPNQTPNSKLVEVN-----PKS 895

Db    814 LRDLQTTDVSLLAIAATLDAIGEKLKDQKARNQQVMDRLCEIEKILGPPKS 864
```

```
RESULT  11
FP1_MYTCO
ID   FP1_MYTCO      STANDARD;      PRT;   872 AA.
AC   Q25434;
DT   30-MAY-2000 (Rel. 39, Created)
DT   30-MAY-2000 (Rel. 39, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ADHESIVE PLAQUE MATRIX PROTEIN PRECURSOR (FOOT PROTEIN 1) (MCFP1).
GN   FP1.
OS   Mytilus coruscus (Sea mussel).
OC   Eukaryota; Metazoa; Mollusca; Bivalvia; Pteriomorphia; Mytiloida;
OC   Mytiloidea; Mytilidae; Mytilus.
OX   NCBI_TaxID=42192;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Foot;
RX   MEDLINE=96394686; PubMed=8798340;
RA   Inoue K., Takeuchi Y., Takeyama S., Yamaha E., Yamazaki F., Odo S.,
RA   Harayama S.;
RT   "Adhesive protein cDNA sequence of the mussel Mytilus coruscus and
RT   its evolutionary implications.";
RL   J. Mol. Evol. 43:348-356(1996).
CC   -!- FUNCTION: PROVIDES ADHESIVENESS TO THE MUSSEL'S FOOT. MUSSELS
CC       PRODUCE ONE OF THE STRONGEST WATER INSOLUBLE GLUE. THE MUSSEL'S
CC       ADHESIVE IS A BUNDLE OF THREADS, CALLED A BYSSUS, FORMED BY A
CC       FIBROUS COLLAGENOUS CORE COATED WITH ADHESIVE PROTEINS.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- TISSUE SPECIFICITY: PRODUCED BY THE BYSSAL GLAND.
CC   -!- DOMAIN: ALMOST EXCLUSIVELY COMPOSED OF REPEATS OF A DECAPEPTIDE.
CC   -!- PTM: PROLINES IN THE REPEAT ARE HYDROXYLATED (SINGLE OR DI-) AND
CC       ALSO TYROSINE (THUS PRODUCING DOPA = 3,4-DIHYDROXYPHENYLALANINE).
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; D63777; BAA09850.1; -.
DR   InterPro; IPR002964; Adhesive_plaq.
DR   InterPro; IPR002965; P_rich_extensn.
DR   PRINTS; PR01216; ADHESIVEI.
DR   PRINTS; PR01217; PRICHEXTENSN.
KW   Signal; Repeat; Hydroxylation.
FT   SIGNAL       1     20       POTENTIAL.
FT   CHAIN       21    872       ADHESIVE PLAQUE MATRIX PROTEIN.
FT   DOMAIN      21     41       NONREPETIVE LINKER.
FT   DOMAIN     124    872       TANDEM REPEATS OF Y-K-[PS]-K-[IP]-[ST]-Y-
FT                               P-[PST]-[ST].
FT   DOMAIN     184    192       NONAPEPTIDE 1.
FT   DOMAIN     213    221       NONAPEPTIDE 2.
SQ   SEQUENCE   872 AA;  101677 MW;  98CC70D7C75FF3C4 CRC64;
```

```
  Query Match           10.3%;  Score 518.5;  DB 1;  Length 872;
  Best Local Similarity  29.2%;  Pred. No. 3.5e-14;
  Matches  276;  Conservative   80;  Mismatches  407;  Indels  181;  Gaps   54;

Qy      19 VVDEAGSGLDNGDFKVTTPDTST--TQH-------NKVSTSPKITTAKPINPRPSLPPNS 69

Db      29 VYGSAYSGASAGAYK-TLPGSHPYGSKHVPVYKPMNKIPT-PYI--SKKSYPAPYKPKGY 84

Qy      70 DTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVL 129

Db      85 YPTKRYQPTYGSKTNYPPIYKPIAKKLSSYKAIKTTYPAYKAKTSYPPSYK-----HKIT 139

Qy     130 AKPTPKAETTTKGPALTTPKEPT-PTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTK 188

Db     140 YPPTYK-------PKITYP--PTYKQKPSYPPSYKPKTTYPPTYK-----PKITYPPTYK 185

Qy     189 SAPT-TPKEPAPT--TTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTT--PKK 243

Db     186 RKPSYTPYKPKATYPPTYKPKITYP----PTYKRKPSYT-----PYKPKTTYPPTYKPKI 236

Qy     244 PAPTTPKEPAPTT----PKEPTPTT--PKEPAPTTKEPAPTTPKEPAPT-APKKPAPTTP 296

Db     237 SYPSIYKPKASYVSSYKSKKTYPPTYKPKISYPPTYKPKPSYP----PTYKPKVTYPPTY 292

Qy     297 KEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPT-----TTKEPAPTTTKSAPTTPK 351

Db     293 K-PKPSYPPTYKPKITYPPT-YKPKPSYPTPYKQKPSYPPIYKSKSSYPTSYKSKKTYPP 350

Qy     352 EPSPTTTKEPA-PTTPKEPAPTTPKKP-APT-TPK-------EPAPTTPKEPAPTTTKKP 401

Db     351 TYKPKITYPPTYKPKPSYPPSYKPKKTYSPTYKPKITYPPTYKPKPSYPPSYKPKTTYPP 410

Qy     402 APTAPKEPAPTTPKETAP-----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEE 456

Db     411 T-YKPKISYPPTYKPKASYVSSYKSKKTYPPTYKPKISYPPTYKPKPSYPPTYKPKITYP 469

Qy     457 PT-PTTPEEPAPTTPKAAAPNTPK-EPAPTTPKEPAPTTP-----KEPAPTT--PKETAP 507
```

```
DR   HSSP; P29768; 1DIL.
DR   InterPro; IPR002860; BNR.
DR   Pfam; PF02012; BNR; 2.
KW   Hydrolase; Glycosidase; Glycoprotein; Repeat; GPI-anchor;
KW   Phosphorylation.
FT   DOMAIN        1    457       CYS-RICH.
FT   REPEAT       23     34       BNR 1.
FT   REPEAT      163    174       BNR 2.
FT   REPEAT      209    220       BNR 3.
FT   DOMAIN      458    588       FIBRONECTIN TYPE-III.
FT   DOMAIN      589   1120       44 X 12 AA TANDEM REPEATS, LTR DOMAIN.
FT   CARBOHYD    342    342       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    394    394       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1125   1125       N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE   1162 AA;  120032 MW;  07049221897C6A40 CRC64;


  Query Match             10.6%;  Score 530.5;  DB 1;  Length 1162;
  Best Local Similarity   29.4%;  Pred. No. 1.5e-14;
  Matches  177;  Conservative  100;  Mismatches  212;  Indels  113;  Gaps   34;

Qy    152 TPTTPKE-PASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTP 210
          ||:|| :  | :||  |  :   || :||  || ::    |:|| : :   |    |:||
Db    599 TPSTPADSSAHSTPSTPVDS---SAHSTPSTPADSSAHGTPSTPVDSSAHGT----PSTP 651

Qy    211 KE-PAPTTTKEPAPTTTKSAPTTPKE-PAPTTPKKP-------APTTPKEPA----PTTP 257
           :  |  |   |  ::  | |:|| :  | :||  |       ||:|| : :    |:||
Db    652 ADSSAHGTPSTPVDSSAHSTPSTPVDSSAHSTPSTPVDSSAHGAPSTPADSSAHGTPSTP 711

Qy    258 KEPTP-TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPS 316
           : :   ||  || ::    |:|| : :       | :||  || ::      |    :
Db    712 VDSSAHGTPSTPADSSAHSTPSTPADSS-------AHSTPSTPADSSAHSTPSTPVDSSA 764

Qy    317 PTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA----PT 372
           ||  || ::  | |:|    || ::    |:|| : |  :|    |:|| : :    |:
Db    765 HGTPSTPADSSAHSTPST---PADSSAHGTPSTPVDSSAHST----PSTPVDSSAHGTPS 817

Qy    373 TP-KKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKL----TP 427
           ||   | :||  |  ::      |   | : |  || ::   | |:||       ||
Db    818 TPVDSSAHSTPSTPVDSSAHGTPSTPVDSSAHSTPSTPADSSAHST-PSTPADSSAHGTP 876

Qy    428 TTP-EKLAPTTPEKPAPTTPEELAPTTPEE----PTPTTPEEPA----PTTP-KAAAPNT 477
           :|| :  | :||  || ::     |:|| :     ||:|| : :    |:||  ::|  |
Db    877 STPVDSSAHSTPSTPADSSAHS-TPSTPVDSSAHSTPSTPADSSAHGTPSTPVDSSAHGT 935

Qy    478 PKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPT 537
           |  || ::    | :||  || ::   | |:||  ::       | :||  |       |
Db    936 PSTPADSS----AHSTPSTPADSSAHST-PSTPADSS-------AHSTPSTPVDSSAHST 983

Qy    538 TTKEPTSTTSDKPAPTTPKGTAPTTPKE-PAPTTPKEPAPTTPKGTAPTTLKEPAPTTPK 596
             || :|  | :|     |:|| :  | :||  || ::   |  | :   | :||
Db    984 -----PSTPADSSAHST-----PSTPADSSAHSTPSTPADSSAHSTPSTPVDSSAHSTPS 1033

Qy    597 KPAPKELAPTTTKGPTSTTSDKPAPTTPK----ETAPTTPKEPAPTTPKKPAPTTPETPP 652
           |    | ::  |  || :|  | :||       :| :||  || ::    |  || ||
Db   1034 TP-----ADSSAHGTPSTPADSSAHSTPSTPVDSSAHSTPSTPADSS----AHGTPSTPA 1084

Qy    653 PTTSEVSTPTTTKEPTTIHKSPDESTP-ELSAEPTPKALENSPKEPGVPTTKTPAATKPE 711
           :::  |||:|  : :: | :|  ||| : ||  ||                  ||| :
Db   1085 DSSAH-STPSTPAD-SSAHGTP--STPADSSAHSTP--------------STPAGSSAN 1125

Qy    712 MT 713
           |
Db   1126 GT 1127


RESULT  10
CPN_DROME
ID   CPN_DROME      STANDARD;      PRT;   865 AA.
AC   Q02910;
DT   01-OCT-1993 (Rel. 27, Created)
DT   01-OCT-1993 (Rel. 27, Last sequence update)
DT   01-FEB-1994 (Rel. 28, Last annotation update)
DE   CALPHOTIN.
GN   CPN OR CAP.
OS   Drosophila melanogaster (Fruit fly).
OC   Eukaryota; Metazoa; Arthropoda; Tracheata; Hexapoda; Insecta;
OC   Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha;
OC   Ephydroidea; Drosophilidae; Drosophila.
OX   NCBI_TaxID=7227;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=CANTON-S;
RX   MEDLINE=93165729; PubMed=8094559;
RA   Martin J.H., Benzer S., Rudnicka M., Miller C.A.;
RT   "Calphotin: a Drosophila photoreceptor cell calcium-binding protein.";
RL   Proc. Natl. Acad. Sci. U.S.A. 90:1531-1535(1993).
RN   [2]
RP   SEQUENCE FROM N.A.
RC   STRAIN=CANTON-S;
RX   MEDLINE=93165730; PubMed=8434015;
RA   Ballinger D.G., Xue N., Harshman K.D.;
RT   "A Drosophila photoreceptor cell-specific protein, calphotin, binds
RT   calcium and contains a leucine zipper.";
RL   Proc. Natl. Acad. Sci. U.S.A. 90:1536-1540(1993).
CC   -!- FUNCTION: MIGHT FUNCTION AS A CALCIUM-SEQUESTERING "SPONGE" TO
CC       REGULATE THE AMOUNT OF FREE CYTOPLASMIC CALCIUM. IT BINDS 0.3 MOL
CC       OF CA+2 PER MOL OF PROTEIN.
CC   -!- SUBUNIT: HOMODIMER (PROBABLE).
CC   -!- SUBCELLULAR LOCATION: CYTOPLASMIC; HYPODENSE COMPARTMENT.
CC   -!- TISSUE SPECIFICITY: SOMA AND AXONS OF PHOTORECEPTOR CELLS OF
CC       COMPOUND EYES AND OCELLI.
CC   -!- DEVELOPMENTAL STAGE: EXPRESSED EARLY IN PHOTORECEPTOR CELL
CC       DEVELOPMENT.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; L02111; AAA28405.1; -.
DR   EMBL; L05080; AAA28420.1; -.
DR   PIR; A47282; A47282.
DR   FlyBase; FBgn0010218; Cpn.
KW   Calcium-binding.
FT   CONFLICT     36     36       A -> AVAPAVVA (IN REF. 2).
FT   CONFLICT     43     43       I -> T (IN REF. 2).
FT   CONFLICT     64     64       I -> V (IN REF. 2).
FT   CONFLICT     76     76       T -> A (IN REF. 2).
FT   CONFLICT    100    100       P -> PP (IN REF. 2).
FT   CONFLICT    126    127       VQ -> AP (IN REF. 2).
FT   CONFLICT    154    154       I -> V (IN REF. 2).
FT   CONFLICT    160    160       S -> T (IN REF. 2).
FT   CONFLICT    534    534       A -> E (IN REF. 2).
FT   CONFLICT    699    699       I -> T (IN REF. 2).
FT   CONFLICT    703    703       V -> L (IN REF. 2).
FT   CONFLICT    721    721       D -> E (IN REF. 2).
SQ   SEQUENCE   865 AA;  84781 MW;  2110417E0B0E7CFE CRC64;


  Query Match             10.6%;  Score 530;  DB 1;  Length 865;
  Best Local Similarity   26.2%;  Pred. No. 1.2e-14;
  Matches  233;  Conservative   73;  Mismatches  433;  Indels  152;  Gaps   41;

Qy    123 APTSKVLAKP-TPK---AETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTT 178
          :| |  :| | ||    |      ||   |       |  | : ||  | ||     | |
Db      8 SPVSAPVAAPVTPSAVAAPVQVVSPAAVAP------APAAPIAVTPVAPPPTLASVQPAT 61

Qy    179 PKEPAPTTTKSAPTTP-KEPAPTTTKEPAPTTPKEPAPTTTKEPAPT----TTKSAPTTP 233
          |||    :|  ||    ||     | |     || :    |          ||| |
Db     62 VTIPAPAPIAAASVTPVASVAPPVVAAPTP-----PAASPVSTPVAVAQIPVAVSAPVAP 116

Qy    234 KEPAPTTPKKPAP-------TTPKEPAPTTPKEPTP-------TTPKEPAPTTKEPAPTT 279
```

```
DR   Pfam; PF00038; filament; 1.
DR   PROSITE; PS00226; IF; 1.
KW   Intermediate filament; Coiled coil; Neurone; Phosphorylation;
KW   Repeat.
FT   DOMAIN        1     97       HEAD.
FT   DOMAIN       98    408       ROD.
FT   DOMAIN      409   1087       TAIL.
FT   DOMAIN      436    517       GLU-RICH (ACIDIC).
FT   DOMAIN      519    886       50 X 6 AA TANDEM REPEATS OF K-S-P-A-E-A.
FT   DOMAIN      887   1087       GLU/LYS-RICH.
FT   DOMAIN       98    129       COIL 1A.
FT   DOMAIN      130    141       LINKER 1.
FT   DOMAIN      142    239       COIL 1B.
FT   DOMAIN      240    261       LINKER 12.
FT   DOMAIN      262    283       COIL 2A.
FT   DOMAIN      284    287       LINKER 2.
FT   DOMAIN      288    408       COIL 2B.
FT   CONFLICT    133    133       K -> QA (IN REF. 2 AND 3).
FT   CONFLICT    199    199       A -> AR (IN REF. 2 AND 3).
FT   CONFLICT    281    281       S -> T (IN REF. 2 AND 3).
FT   CONFLICT    492    492       L -> G (IN REF. 2 AND 3).
FT   CONFLICT    551    551       P -> PREAKSP (IN REF. 3).
FT   CONFLICT    689    712       MISSING (IN REF. 3).
FT   CONFLICT    714    714       G -> A (IN REF. 3).
FT   CONFLICT    814    814       V -> M (IN REF. 2 AND 3).
FT   CONFLICT    843    843       T -> N (IN REF. 2 AND 3).
SQ   SEQUENCE   1087 AA;  116612 MW;  57BAC76A38ED1CB9 CRC64;


  Query Match           10.6%;  Score 533;  DB 1;  Length 1087;
  Best Local Similarity  27.5%;  Pred. No. 1.1e-14;
  Matches  216;  Conservative   81;  Mismatches  330;  Indels  158;  Gaps   38;

Qy      21 DEAGSGLDNGDFKVT-----TPDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTSKET 75
           :|   |     | :|     |  ||  | || :   |   :              : |||
Db     404 EECRIGFGPSPFSLTEGLPKIPSIST--HIKVKSEEMIKVVE-----------KSEKET 449

Qy      76 SLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPK 135
           :  :   :  |  |  :    |:|   : :  :  |:  |   :| ::  | |  :
Db     450 VIVEGQTEEIRVTEGVTEEEDKEAQGQEGEEAEEGEEKEEEELAAATSPPAEEAASPEKE 509

Qy     136 AETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPK 195
           ::  |  | :  :  :|   | ||    :  :|   || |   | |    : : | :|
Db     510 TKSRVKEEAKSPGEAKSPGEAKSPA----EAKSPGEAKS-PGEAKSPGEAKSPAEPKSPA 564

Qy     196 EP-APTTTKEPAPTTPKEPAPTTTKEP----APTTTKSAPTTPKEPA----PTTPKKPA- 245
           || :|   | ||   || ||  | | |    :|:  || |   | ||    |   | ||
Db     565 EPKSPAEAKSPA--EPKSPA--TVKSPGEAKSPSEAKS-PAEAKSPAEAKSPAEAKSPAE 619

Qy     246 ---PTTPKEPA----PTTPKEP----TPTTPKEPAPTTKEPAPTTPKEPA----PTAPKK 290
              |  | ||    | | | |     :|:  | ||   :  :|   | ||    |   |
Db     620 AKSPAEAKSPAEAKSPATVKSPGEAKSPSEAKSPA---EAKSPAEAKSPAEAKSPAEVKS 676

Qy     291 P----APTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPA----PTTTKSAPTTTKEPAPTT 342
           |     :|  || ||    | ||    :  ||   | ||    |   || |   | ||
Db     677 PGEAKSPAEPKSPAEA--KSPA----EVKSPAEAKSPAEVKSPGEAKS-PAAVKSPAEAK 729

Qy     343 TKSAPTTPKE-PSPTTTKEPA----PTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTT 397
           : :|  :| | |  ||   | ||    |   | |      : | |  || | :  :  :|
Db     730 SPAAVKSPGEAKSPGEAKSPAEAKSPAEAKSPIEVKSPEKAKTPVKEGAKSPAEAKSPEK 789

Qy     398 TKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEP 457
           | |     | ||    | |  :| |     ||| |   : |   || :   | | :
Db     790 AKSPVKEDIKPPAEAKSPEKA-KSPVKEGAKPPEKAKPLDVKSPEAQTPVQEEATVPTDI 848

Qy     458 TPTTPEE-PAPTTPKAAAPNTPKEPAPTT----PKEPAPTTP-------KEPAPTTPKET 505
           |  ||:  :|   || :|   || | |:    ||:   :|        |||    :|
Db     849 RP--PEQVKSPAKEKAKSPE--KEEAKTSEKVAPKKEEVKSPVKEEVKAKEPPKKVEEEK 904

Qy     506 APTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSD--KPAPTTPKGTAPTTP 563
           |||  |  : |: ||    ||  :|   | |::|  :|::  |      |    :
Db     905 TLPTPKTEAKESKKDEAPKEAPKPKVEEKKETPTEKPKDSTAEAKKEEAGEKKKAVASEE 964

Qy     564 KEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTT 623
           : ||    || |   ||    ||  |   |  |:|:     || |: |
Db     965 ETPAKLGVKEEA--KPKEKTETTKTEAEDTKAKEPS----KPTETEKP----------- 1006

Qy     624 PKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSA 683
                 ||  |  |:|                    | :| ||  : |:|
Db    1007 --------KKEEMPAAPEK----------------KDTKEEKTTESRKPEEK------ 1034

Qy     684 EPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETA 743
              || :|   ||     :| |  :||:     |  :|::     ||     | |:: |
Db    1035 ---PK-MEAKVKEDDKSLSKEP--SKPKTEKAEKSSSTDQKESQPPE------KTTEDKA 1082

Qy     744 TTTEK 748
           |  ||
Db    1083 TKGEK 1087


RESULT   9
TCNA_TRYCR
ID   TCNA_TRYCR     STANDARD;      PRT;  1162 AA.
AC   P23253;
DT   01-NOV-1991 (Rel. 20, Created)
DT   01-NOV-1991 (Rel. 20, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   SIALIDASE (EC 3.2.1.18) (NEURAMINIDASE) (NA) (MAJOR SURFACE ANTIGEN).
GN   TCNA.
OS   Trypanosoma cruzi.
OC   Eukaryota; Euglenozoa; Kinetoplastida; Trypanosomatidae; Trypanosoma.
OX   NCBI_TaxID=5693;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=SILVIO X-10/4;
RX   MEDLINE=91277609; PubMed=1711561;
RA   Pereira M.E.A., Mejia J.S., Ortega-Barria E.,. Matzilevich D.,
RA   Prioli R.P.;
RT   "The Trypanosoma cruzi neuraminidase contains sequences similar to
RT   bacterial neuraminidases, YWTD repeats of the low density lipoprotein
RT   receptor, and type III modules of fibronectin.";
RL   J. Exp. Med. 174:179-191(1991).
RN   [2]
RP   SUBCELLULAR LOCATION.
RX   MEDLINE=91376547; PubMed=1896773;
RA   Prioli R.P., Mejia J.S., Aji T., Aikawa M., Pereira M.E.A.;
RT   "Trypanosoma cruzi: localization of neuraminidase on the surface of
RT   trypomastigotes.";
RL   Trop. Med. Parasitol. 42:146-150(1991).
CC   -!- FUNCTION: DEVELOPMENTALLY REGULATED NEURAMINIDASE IMPLICATED IN
CC       PARASITE INVASION OF CELLS.
CC   -!- CATALYTIC ACTIVITY: HYDROLYSIS OF 2,3-, 2,6- AND 2,8-GLYCOSIDIC
CC       LINKAGES JOINING TERMINAL NON-REDUCING N- OR O-ACYLNEURAMINYL
CC       RESIDUES TO GALACTOSE, N-ACETYLHEXOSAMINE, OR N- OR O-ACYLATED
CC       NEURAMINYL RESIDUES IN OLIGOSACCHARIDES, GLYCOPROTEINS,
CC       GLYCOLIPIDS OR COLOMINIC ACID.
CC   -!- SUBCELLULAR LOCATION: ATTACHED TO THE MEMBRANE BY A GPI-ANCHOR
CC       (POSSIBLE).
CC   -!- DEVELOPMENTAL STAGE: MAXIMAL ACTIVITY IN TRYPOMASTIGOTES, MINIMUM
CC       IN EPIMASTIGOTES AND NO DETECTABLE IN AMASTIGOTES.
CC   -!- MISCELLANEOUS: THE VARIABLE LENGTHS OF THE LONG TANDEM REPEAT
CC       DOMAIN COULD ACCOUNT IN PART FOR THE POLYMORPHISM OF THE TNCA
CC       PROTEIN.
CC   -!- SIMILARITY: BELONGS TO FAMILY 33 OF GLYCOSYL HYDROLASES.
CC   -!- SIMILARITY: CONTAINS 3 BNR REPEATS.
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; M61732; AAA30255.1; -.
DR   PIR; JH0557; JH0557.
```

```
CC   -!- PTM: EXTENSINS CONTAIN A CHARACTERISTIC REPEAT OF THE PENTAPEPTIDE
CC       SER-PRO(4). THE PROLINE RESIDUE IS HYDROXYLATED AND THEN
CC       GLYCOSYLATED.
CC   ---------------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------------
DR   EMBL; X13885; CAA32090.1; -.
DR   PIR; S06733; S06733.
KW   Repeat; Cell wall; Glycoprotein; Signal; Structural protein;
KW   Hydroxylation.
FT   SIGNAL        1      ?
FT   CHAIN         ?    620       EXTENSIN.
FT   REPEAT       70     73        H-A-P-P.
FT   REPEAT      148    151        H-A-P-P.
FT   DOMAIN      229    242        2 X 7 AA TANDEM REPEATS OF T-H-R-H-A-P-P.
FT   REPEAT      229    235        1.
FT   REPEAT      236    242        2.
FT   DOMAIN      205    620        CONTAINS THE SER-PRO(4) REPEATS.
FT   DOMAIN      499    600        3 X APPROXIMATE TANDEM REPEATS.
SQ   SEQUENCE   620 AA;  65406 MW;  641DD2278AB28524 CRC64;


  Query Match             11.0%;  Score 551;  DB 1;  Length 620;
  Best Local Similarity   27.2%;  Pred. No. 1.5e-15;
  Matches  170;  Conservative   64;  Mismatches  304;  Indels   88;  Gaps   24;

Qy    136 AETTTK-GPALTTP--KEPTPTT-----PKEPASTTPKE---PTPTTIKSAPTTPKEPAP 184
          || ||: |  |  |   :| |::     |  | :| |     |:|   : ||     | |
Db     24 AEATTQYGGYLPPPVTSQPPPSSIGLSPPSAPTTTPPSRGHVPSP---RHAPPRHAYPPP 80

Qy    185 TTTKSAPTTPKEP------APTTTKEPAPT---TPKEPAPTTTKEPAPTTTKSAPTTPKE 235
          :    |:    |       |:    | |:   :|  | |:    |        |:  :
Db     81 SHGHLPPSVGGPPPHRGHLPPSRGFNPPPSPVISPSHPPPSYGAPPPSHGPGHLPSHGQR 140

Qy    236 PAPTTPKKPAPT---TPKEPAPTTPKEPTPTTPKEPAPTTK--EPAPTTPKEPAPTAPKK 290
          |      |:|:    |     | |:   | : : |:| ::   | | |  :| ||
Db    141 P-------PSPSHGHAPPSGGHTPPRGQHPPSHRRPSPPSRHGHPPPPTYAQPPPTPIYS 193

Qy    291 PAPTTPKEPAPTTPKEPAPTTTKEPSPTTPK---EPAPTTTKSAPTTTKEPAPTTTKS-- 345
          |:|    :| ||    | | |  :|:|: |    :| | | : || | :   ||   |
Db    194 PSPQV--QPPPT--YSPPPPTHVQPTPSPPSRGHQPQPPTHRHAPPTHRHAPPTHQPSPL 249

Qy    346 ---APTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPT---TPKEPAPTTTK 399
              |:  ::| | |   | |   : | |:    | | |   | |:   :|  || : :
Db    250 RHLPPSPRRQPQPPTYSPPPPAYAQSPQPSPTYSPPPPTYSPPPPSPIYSPPPPAYSPSP 309

Qy    400 KPAPT---APKEPAPTTPKETAPTTPKKL-TPTT------PEKLAPTTPEKPAPTTPEEL 449
          | ||   :|  || : |   :|  | |  |::      |   :|  |   :|  | |
Db    310 PPTPTPTFSPPPPAYSPPPTYSPPPPTYLPLPSSPIYSPPPPVYSPPPPPSYSPPPPTYL 369

Qy    450 APTTPEEPTPTTPEEPAPTTPKAAAP---NTPKEPAPTTPKEPAPTTPKEPAPTTPKETA 506
          | |  | | :   | ||  ::  |    :|  ||| |   | ||     | | |  :
Db    370 PPPPPSSPPPPSFSPPPPTYEQSPPPPPAYSPPLPAPPTYSPPPPT--YSPPPPTYAQPP 427

Qy    507 PTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEP 566
          |      | |   | |     | |     | |  |       :| |  |  : |     |
Db    428 P------LPPTYSPPPPAYSPPPPPTYSPPPPTYSPPPPAYAQPPPPPPTYSPPPPAYSP 481

Qy    567 APTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKE 626
          | :|    |       | |   | |     | |    |   : || |    |:| |
Db    482 PPPSPIYSPPPPQVQPLPPTFSPPPPRRIHLPPPPHRQP---RPPTPTYGQPPSPPTFSP 538

Qy    627 TAPTTPKEPAPT--TPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAE 684
          .|     | |    |: | ||  : | | |     |     |   |:  |   ||
Db    539 PPPRQIHSPPPPHWQPRTPTPTYGQPPSPPTFSAPPPRQIHSPPPPHRQPRPPTPTYGQP 598

Qy    685 PTPKALENSPKEPGVPTTKTPAATKP 710
```

```
          |:|            ||| :| :  |
Db    599 PSP-----------PTTYSPPSPPP 612


RESULT   8
NFH_MOUSE
ID   NFH_MOUSE      STANDARD;      PRT;  1087 AA.
AC   P19246; Q61959;
DT   01-NOV-1990 (Rel. 16, Created)
DT   01-AUG-1992 (Rel. 23, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   NEUROFILAMENT TRIPLET H PROTEIN (200 KDA NEUROFILAMENT PROTEIN)
DE   (NEUROFILAMENT HEAVY POLYPEPTIDE) (NF-H).
GN   NEFH OR NFH.
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=89121513; PubMed=3220257;
RA   Julien J.-P., Cote F., Beaudet L., Sidky M., Flavell D., Grosveld F.,
RA   Mushynski W.;
RT   "Sequence and structure of the mouse gene coding for the largest
RT   neurofilament subunit.";
RL   Gene 68:307-314(1988).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=89089138; PubMed=3145094;
RA   Shneidman P.S., Carden M.J., Lees J.F., Lazzarini R.A.;
RT   "The structure of the largest murine neurofilament protein (NF-H) as
RT   revealed by cDNA and genomic sequences.";
RL   Brain Res. 464:217-231(1988).
RN   [3]
RP   SEQUENCE FROM N.A.
RC   STRAIN=SWISS WEBSTER; TISSUE=Brain;
RA   Carden M.J.;
RL   Submitted (MAR-1994) to the EMBL/GenBank/DDBJ databases.
CC   -!- FUNCTION: NEUROFILAMENTS USUALLY CONTAIN THREE IF PROTEINS: L, M,
CC       AND H WHICH ARE INVOLVED IN THE MAINTENANCE OF NEURONAL CALIBER.
CC       NF-H HAS AN IMPORTANT FUNCTION IN MATURE AXONS THAT IS NOT
CC       SUBSERVED BY THE TWO SMALLER NF PROTEINS.
CC   -!- PTM: THERE ARE A NUMBER OF REPEATS OF THE TRIPEPTIDE K-S-P, NFH IS
CC       PHOSPHORYLATED ON A NUMBER OF THE SERINES IN THIS MOTIF. IT IS
CC       THOUGHT THAT PHOSPHORYLATION OF NFH RESULTS IN THE FORMATION OF
CC       INTERFILAMENT CROSS BRIDGES THAT ARE IMPORTANT IN THE MAINTENANCE
CC       OF AXONAL CALIBER.
CC   -!- PTM: PHOSPHORYLATION SEEMS TO PLAY A MAJOR ROLE IN THE FUNCTIONING
CC       OF THE LARGER NEUROFILAMENT POLYPEPTIDES (NF-M AND NF-H), THE
CC       LEVELS OF PHOSPHORYLATION BEING ALTERED DEVELOPMENTALLY AND
CC       COINCIDENT WITH A CHANGE IN THE NEUROFILAMENT FUNCTION.
CC   -!- SIMILARITY: BELONGS TO THE INTERMEDIATE FILAMENT FAMILY.
CC   -!- CAUTION: REF.2 SEQUENCE DIFFERS FROM THAT SHOWN IN POSITIONS 534
CC       TO 716 AND IS SHORTER DUE TO FRAMESHIFTS.
CC   ---------------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------------
DR   EMBL; M24496; AAA39813.1; -.
DR   EMBL; M23349; AAA39813.1; JOINED.
DR   EMBL; M24494; AAA39813.1; JOINED.
DR   EMBL; M24495; AAA39813.1; JOINED.
DR   EMBL; M35131; AAA39809.1; ALT_FRAME.
DR   EMBL; Z31012; CAA83229.1; -.
DR   PIR; JT0368; QFMSH.
DR   PIR; A43778; A43778.
DR   MGD; MGI:97309; Nfh.
DR   InterPro; IPR001664; IF.
```

```
CC       APICAL REGION OF THE SPERM HEAD (BY SIMILARITY).
CC   -!- TISSUE SPECIFICITY: IN TESTIS; PRIMARILY IN HAPLOID SPERMATIDS.
CC   -!- DOMAIN: THE MAM DOMAINS PROBABLY MEDIATE SPERM ADHESION TO THE
CC       ZONA PELLUCIDA.
CC   -!- DOMAIN: DURING SPERM MIGRATION THROUGH THE REPRODUCTIVE TRACTS,
CC       THE MUCIN-LIKE DOMAIN MIGHT INHIBIT INAPPROPRIATE TRAPPING OF
CC       SPERMATOZOA OR PROMOTING ADHESION TO THE OVIDUCTAL ISTHMUS.
CC   -!- DOMAIN: THE VWFD DOMAINS 2 AND 3 MAY MEDIATE COVALENT
CC       OLIGOMERIZATION (BY SIMILARITY TO HUMAN INTESTINAL MUCIN MUC2).
CC   -!- SIMILARITY: CONTAINS 3.5 MAM DOMAINS.
CC   -!- SIMILARITY: CONTAINS 4.5 VWFD DOMAINS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; AF053356; AAC78790.1; -.
DR   EMBL; U83191; AAC51208.1; -.
DR   MIM; 602372; -.
DR   InterPro; IPR000561; EGF-like.
DR   InterPro; IPR000998; MAM.
DR   InterPro; IPR002965; P_rich_extensn.
DR   InterPro; IPR002919; TIL.
DR   InterPro; IPR003328; TILa.
DR   InterPro; IPR001007; VWFC.
DR   InterPro; IPR001846; Vwd.
DR   Pfam; PF00629; MAM; 3.
DR   Pfam; PF01826; TIL; 4.
DR   Pfam; PF02345; TILa; 4.
DR   Pfam; PF00094; vwd; 4.
DR   PRINTS; PR01217; PRICHEXTENSN.
DR   SMART; SM00137; MAM; 2.
DR   SMART; SM00214; VWC; 1.
DR   SMART; SM00011; VWC_def; 3.
DR   SMART; SM00216; VWD; 3.
DR   PROSITE; PS01186; EGF_2; 3.
DR   PROSITE; PS00740; MAM_1; 1.
DR   PROSITE; PS50060; MAM_2; 4.
KW   Glycoprotein; Transmembrane; Cell adhesion; Repeat.
FT   NON_TER       1      1
FT   DOMAIN       <1    109       MAM 1.
FT   DOMAIN      112    136       MAM 2 (PARTIAL).
FT   DOMAIN      161    326       MAM 3.
FT   DOMAIN      322    446       MAM 4.
FT   DOMAIN      483    951       66 X HEPTAPEPTIDE REPEATS (APPROXIMATE)
FT                                (MUCIN-LIKE DOMAIN).
FT   DOMAIN      953   1065       VWFD 1 (PARTIAL).
FT   DOMAIN     1066   1454       VWFD 2.
FT   DOMAIN     1455   1861       VWFD 3.
FT   DOMAIN     1862   2292       VWFD 4.
FT   DOMAIN     2293   2684       VWFD 5.
FT   DOMAIN        ?      ?       EGF-LIKE.
FT   CARBOHYD     74     74       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    403    403       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1023   1023       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1099   1099       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1618   1618       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1737   1737       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1832   1832       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1878   1878       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   2136   2136       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   2505   2505       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CONFLICT   2374   2379       NNQKMA -> RAGPGP (IN REF. 1).
FT   NON_TER    2700   2700
SQ   SEQUENCE   2700 AA;  293013 MW;  80E60CC0B12277B1 CRC64;
```

```
  Query Match           11.1%;  Score 555.5;  DB 1;  Length 2700;
  Best Local Similarity 32.6%;  Pred. No. 3e-15;

  Matches  199;  Conservative   64;  Mismatches  240;  Indels  107;  Gaps   32;

Qy      58 PINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKT 117
           |:   | ||| |  |      |   |||  |:  | : |                | | ||||
Db     445 PVKVLPELPPVSPVSS----TGPSETTGLTENPTISTK------------KPTVSIEKP 487

Qy     118 SAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPT-----PTTI 172
           |     ||       |  |     |: | ::| |   || | |  :   ::||     || :
Db     488 SVTTEKPT-----VPKEKPTIPTEKPTISTEK---PTIPSEKPNMPSEKPTIPSEKPTIL 539

Qy     173 KSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPA-PTTTKSAPT 231
             || | |  ||      ||   |      |:|  |||| |   |  :||  ||   | ||
Db     540 TEKPTIPSE-KPTIPSEKPTISTEKPTVPTEE--PTTPTEETTTYMEEPVIPTEKPSIPT 596

Qy     232 TPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKP 291
             ::|:  | |   ||   |    : ::||    |   || |   || | | :  :|:|
Db     597 --EKPSIPTEK---PTISMEETIISTEKPTICPEKPTIPTEK---PTIPTEKSTISPEK- 647

Qy     292 APTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPK 351
           |||| |  || | |     :|::|:  | | |  :  |    | |  ||     || |
Db     648 -PTTPTE-KPTIPTEKPTISTEKPTIPTEK-PTISPEKLTIPTEKLTIPT---EKPTIPT 701

Qy     352 EPSPTTTKEPAPTTPKEPAPTTPKKPA-----PTTPKEPAPTTPKEPAPTTTKKPAPTAP 406
           |    :|:|  |||| |    : :||:     || | |   |: :|   :| |   |
Db     702 EKPTISTEE--PTTPTEETTISTEKPSIPMEKPTLPTEETTTSVEETTISTEKLTIPM-- 757

Qy     407 KEPAPTTPKETAPT-----TPKKLTPTTPEKLA-----PTTPEKPAPTTPEELAPTTPEE 456
           ::|  :| | | ||     :|:|||  | |||      || | :    : |:|  | | |
Db     758 EKPTISTEKPTIPTEKPTISPEKLTIPT-EKLTIPTEKPTIPIEETTISTEKL--TIPTE 814

Qy     457 PTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPT 516
           :||:|  :| |           || | |  || | |        ||  :| | | ||
Db     815 KPTISPEKPTISTEK----------PTIPTE-KPTIPTE------ETTISTEKLTIPT 855

Qy     517 TLKEPAPTTPKKPAPKELAPTTTKEPTSTT------SDKPAPTTPKGTAPTTPKEPAPTT 570
             ::|  :  |   | |    :|::||  |       ::|   | | | ||  ::
Db     856 --EKPTISPEKLTIPTEKPTISTEKPTIPTEKLTIPTEKPTIPTEKPTIPT--EKLTALR 911

Qy     571 PKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKET--A 628
           |  |:||   | |   :   ||:||       :  |||   :||    | |     |:
Db     912 PPHPSPTA-TGLAALVMSPHAPSTPMTSV---ILGTTTTSRSSTGMSCP-PNARYESCAC 966

Qy     629 PTTPKEPAPT 638
           | : | | |:
Db     967 PASCKSPRPS 976
```

```
RESULT   7
EXTN_TOBAC
ID   EXTN_TOBAC     STANDARD;      PRT;   620 AA.
AC   P13983;
DT   01-JAN-1990 (Rel. 13, Created)
DT   01-JAN-1990 (Rel. 13, Last sequence update)
DT   01-MAR-1992 (Rel. 21, Last annotation update)
DE   EXTENSIN PRECURSOR (CELL WALL HYDROXYPROLINE-RICH GLYCOPROTEIN).
GN   HRGPNT3.
OS   Nicotiana tabacum (Common tobacco).
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC   Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots;
OC   Asteridae; euasterids I; Solanales; Solanaceae; Nicotiana.
OX   NCBI_TaxID=4097;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=CV. XANTHI; TISSUE=Leaf;
RX   MEDLINE=90128263; PubMed=2612909;
RA   Keller B., Lamb C.J.;
RT   "Specific expression of a novel cell wall hydroxyproline-rich
RT   glycoprotein gene in lateral root initiation.";
RL   Genes Dev. 3:1639-1646(1989).
CC   -!- FUNCTION: HAS A SPECIALIZED STRUCTURAL FUNCTION, POSSIBLY IN
CC       THE MECHANICAL PENETRATION OF THE CORTEX AND EPIDERMIS OF THE
CC       MAIN ROOT.
CC   -!- SUBCELLULAR LOCATION: EXTRACELLULAR MATRIX.
```

```
RT   protein.";
RL   Biotechnol. Prog. 6:171-177(1990).
RN   [2]
RP   PARTIAL SEQUENCE, AND POST-TRANSLATIONAL MODIFICATIONS.
RX   MEDLINE=83135732; PubMed=6298211;
RA   Waite J.H.;
RT   "Evidence for a repeating 3,4-dihydroxyphenylalanine- and
RT   hydroxyproline-containing decapeptide in the adhesive protein of the
RT   mussel, Mytilus edulis L.";
RL   J. Biol. Chem. 258:2911-2915(1983).
CC   -!- FUNCTION: PROVIDES ADHESIVENESS TO THE MUSSEL'S FOOT. MUSSELS
CC       PRODUCE ONE OF THE STRONGEST WATER INSOLUBLE GLUE. THE MUSSEL'S
CC       ADHESIVE IS A BUNDLE OF THREADS, CALLED A BYSSUS, FORMED BY A
CC       FIBROUS COLLAGENOUS CORE COATED WITH ADHESIVE PROTEINS.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- TISSUE SPECIFICITY: PRODUCED BY THE BYSSAL GLAND.
CC   -!- DOMAIN: ALMOST EXCLUSIVELY COMPOSED OF REPEATS OF A DECAPEPTIDE.
CC   -!- PTM: THE DECAPEPTIDE A-K-P-S-Y-P-P-T-Y-K IS POST-TRANSLATIONALLY
CC       MODIFIED AS FOLLOWING: THE SIXTH AND SEVENTH RESIDUES ARE
CC       HYDROXYLATED AND THE PENULTIMATE IS A 3,4-DIHYDROXYPHENYLALANINE
CC       (DOPA) DERIVED FROM TYROSINE.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X54422; CAA38294.1; -.
DR   InterPro; IPR002964; Adhesive_plaq.
DR   InterPro; IPR002965; P_rich_extensn.
DR   PRINTS; PR01216; ADHESIVEI.
DR   PRINTS; PR01217; PRICHEXTENSN.
KW   Repeat; Hydroxylation.
FT   NON_TER       1      1
FT   DOMAIN       67    870       TANDEM REPEATS OF Y-K-[PA]-K-[LP]-[ST]-Y-
FT                                 P-P-[ST].
SQ   SEQUENCE   875 AA;  100412 MW;  6EA85312748CAACE CRC64;
```

Query Match 11.1%; Score 556.5; DB 1; Length 875;
Best Local Similarity 28.7%; Pred. No. 1.2e-15;
Matches 272; Conservative 106; Mismatches 375; Indels 195; Gaps 56;

```
Qy     42 TQHNKVSTSPKITTAKPINP------------RP--SLPPNSDTSKETSLTVNKETTVET 87
          |:|  |    | : : |  |              ||  | ||    ||   | : |: :
Db      1 TKHEPVYKPVKTSYSAPYKPPTYQPLKKKVDYRPTKSYPPTYG-SKTNYLPLAKKLSSYK 59

Qy     88 KETTTTNKQTSTDG--KEKTT---SAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKG 142
             || | :|:     | | |   : |   |   |          |:   || ||: :
Db     60 PIKTTYNAKTNYPPVYKPKMTYPPTYKPKPSYPPTYKSKPTYKPKITYPPTYKAKPSY-- 117

Qy    143 PALTTPKEPTPTTPKE----PASTTPKEPTPTTIKSAPTTP-----KEPAPTTTKSAPTT 193
          |:   ||:  | | |     | :  ||   | | |  |: |     |:  |:: |: |:
Db    118 PSSYKPKKTYPPTYKPKLTYPPTYKPKPSYPPTYKPKPSYPPSYKTKKTYPSSYKAKPSY 177

Qy    194 PKEPAPTTTKEPA--PTTPKEPA--------PTTTKEPA-PTTTKSAPTTPK--EPAPTT 240
          |    ||   :|:  ||   :|:         ||   :|  |:| |: |: |   :  ||
Db    178 P----PTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTYPSTYKAKPSYPPTYKAKPTY 233

Qy    241 PKKPA--PTTPKEPA--PTTPKEPT--PTTPKEP---APTTKEPAPTTPKEPA--PTAPK 289
            ||:  ||   :|:  ||   :|:  ||   :|   |  | :  ||   :|:  ||
Db    234 KAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTYKAKPTYKAKPSYPPTYKA 293

Qy    290 KPA--PTTPKEPA--PTTPKEPA--PTTTKEPS--PTTPKEPA-PTTTKSAPTTTKEPA- 339
          ||:  ||   :|:  ||   :|:  ||   :||  ||   :|: | | |: ||   :|
Db    294 KPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTY 353

Qy    340 PTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPA--PTTPKEPA--PTTPKEPAP 395
          |:| |: |: |    ||   :|:     :  ||   ||:  ||   :|:  ||   :|:
Db    354 PSTYKAKPSYP----PTYKAKPSYPPTYKAKPTYKAKPSYPPTYKAKPSYPPTYKAKPSY 409

Qy    396 TTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPE 455
            | |  ||   :|       | |:|  |  |: |    |:   ||:     :  ||
Db    410 PPTYKAKPTYKAKP-------TYPST-YKAKPSYP----PSYKAKPSYPPTYKAKPTYKA 457

Qy    456 EPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKE----PAPTTPKETAPTT-- 509
          :||       |:| ||     |  ||    |  | | |      |:  ||:| | |
Db    458 KPT-------YPSTYKA----KPSYPASYKAKPSYPPTYKSKSSYPSSYKPKKTYPPTYK 506

Qy    510 PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTT----PKE 565
          || |   | | | |: |    ||   | | |              || : | |    |
Db    507 PKLTYKPTYK-PKPSYPPSYKPKTTYPPTYK-------------PKISYPPTYKAKPSY 551

Qy    566 PAPTTPKEPAPTTPKGT---APTTLKEPA--PTTPKKPA--PKELAPTTTKG-PTSTTSD 617
          ||    |   | | |       ||   :|:  ||   ||:   |   |:| |  |:   :
Db    552 PATYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYKAKPTYPSTYKAKPSYPPTY 611

Qy    618 KPAPTTPKETAPTTPKEPA--PTTPKKPA-PTT----PETPPPTTSEVSTPTTTKE---- 666
          |  |: |    ||   :|:  ||   ||  |:|    |  ||    ::| | | |
Db    612 KAKPSYP----PTYKAKPSYPPTYKAKPTYPSTYKAKPSYPPTYKPKISYPPTYKAKPSY 667

Qy    667 PTTIHKSPDESTPELSAEPTPKALENSPK----EPGVPTT--KTPA-----ATKPEMTTT 715
          | | :|:      |   |:|| ||   :|     :|  | |    |:       ||    |
Db    668 PPT-YKAKPSYPPTYKAKPTYKAKPTNPSTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPT 726

Qy    716 AKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKIT 775
            | | | :   | | |  | |  | :   |   | ::|        :   |    |
Db    727 YKAKPTYKAKPTYPSTYKAKP--TYKAKPTYPPTYKAK-------PSYPPTYKPKPSYPP 777

Qy    776 TLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKK---P 832
          | |: :: |     |||   |    ||: | |  |  :        |||  |     |   |
Db    778 TYKSKSIYPSSYKPKKTYPPT---YKPKLTYPPTYK--------PKPSYPPSYKPKITYP 826

Qy    833 TSTKKPKTMPRVRKPKTTPTP------------RKMTSTMPELNPTSR 868
          :: |   : |   | ||:  |              :  ||  |   ||:|
Db    827 STYKLKPSYPPTYKSKTSYPPTYNKKISYPSSYKAKTSYPPAYKPTNR 874
```

```
RESULT   6
ZAN_HUMAN
ID   ZAN_HUMAN      STANDARD;      PRT;  2700 AA.
AC   Q9Y493; O00218;
DT   20-AUG-2001 (Rel. 40, Created)
DT   20-AUG-2001 (Rel. 40, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ZONADHESIN (FRAGMENT).
GN   ZAN.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE OF 1-2379 FROM N.A.
RX   MEDLINE=99018118; PubMed=9799793;
RA   Glockner G., Scherer S., Schattevoy R., Boright A., Weber J.,
RA   Tsui L.C., Rosenthal A.;
RT   "Large-scale sequencing of two regions in human chromosome 7q22:
RT   analysis of 650 kb of genomic sequence around the EPO and CUTL1 loci
RT   reveals 17 genes.";
RL   Genome Res. 8:1060-1073(1998).
RN   [2]
RP   SEQUENCE OF 2338-2700 FROM N.A.
RC   TISSUE=Testis;
RX   MEDLINE=97271566; PubMed=9126492;
RA   Gao Z., Harumi T., Garbers D.L.;
RT   "Chromosome localization of the mouse zonadhesin gene and the human
RT   zonadhesin gene (ZAN).";
RL   Genomics 41:119-122(1997).
CC   -!- FUNCTION: BINDS IN A SPECIES-SPECIFIC MANNER TO THE ZONA PELLUCIDA
CC       OF THE EGG. MAY BE INVOLVED IN GAMETE RECOGNITION AND/OR
CC       SIGNALING.
CC   -!- SUBUNIT: PROBABLY FORMS COVALENT OLIGOMERS.
CC   -!- SUBCELLULAR LOCATION: TYPE I MEMBRANE PROTEIN, EXCLUSIVELY ON THE
```

```
DR   EMBL; M61170; AAB53150.1; -.
DR   EMBL; X52229; CAA36478.1; ALT_SEQ.
DR   EMBL; X52228; CAA36477.1; ALT_SEQ.
DR   EMBL; M35093; AAB59612.1; ALT_SEQ.
DR   EMBL; Z17324; CAA78972.1; -.
DR   EMBL; Z17325; CAA78973.1; -.
DR   EMBL; M31823; AAA35757.1; -.
DR   EMBL; S81781; AAD14376.1; ALT_INIT.
DR   EMBL; S81736; AAD14369.1; ALT_INIT.
DR   EMBL; M21868; AAA59874.1; ALT_SEQ.
DR   PIR; A35175; A35175.
DR   PIR; B35175; B35175.
DR   PIR; S10218; S10218.
DR   GlycoSuiteDB; P15941; -.
DR   MIM; 158340; -.
DR   MIM; 113720; -.
DR   InterPro; IPR000082; SEA.
DR   Pfam; PF01390; SEA; 1.
DR   SMART; SM00200; SEA; 1.
DR   PROSITE; PS50024; SEA; 1.
KW   Glycoprotein; Signal; Cytoskeleton; Actin-binding; Transmembrane;
KW   Repeat; Alternative splicing.
FT   SIGNAL        1     23       POTENTIAL.
FT   CHAIN        24   1255       MUCIN 1.
FT   DOMAIN       24   1162       EXTRACELLULAR (POTENTIAL).
FT   TRANSMEM   1163   1186       POTENTIAL.
FT   DOMAIN     1187   1255       CYTOPLASMIC (POTENTIAL).
FT   DOMAIN       81    960       44 X 20 AA TANDEM REPEATS.
FT   DOMAIN     1034   1151       SEA.
FT   CARBOHYD    957    957       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    975    975       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1029   1029       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1055   1055       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1133   1133       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   VARSPLIC     19     19       T -> TATTAPKPAT (IN ISOFORM B).
FT   VARSPLIC     20     22       MISSING (IN ISOFORM C).
FT   VARSPLIC     20     31       MISSING (IN ISOFORM D).
FT   VARSPLIC    126    905       MISSING (IN POLYMORPHIC EPITHELIAL
FT                                ISOFORM).
FT   VARSPLIC   1077   1087       FLQIYKQGGFL -> VSIGLSFPMLP (IN SECRETED
FT                                ISOFORM).
FT   VARSPLIC   1088   1255       MISSING (IN SECRETED ISOFORM).
FT   CONFLICT      2      2       T -> A (IN REF. 11).
FT   CONFLICT    134    134       P -> Q (IN REF. 9).
FT   CONFLICT    154    154       P -> Q (IN REF. 9).
FT   CONFLICT   1021   1021       S -> T (IN REF. 3).
FT   CONFLICT   1251   1251       A -> T (IN REF. 3).
SQ   SEQUENCE   1255 AA;  122072 MW;  5E28DFC4DE7D9A82 CRC64;
```

```
  Query Match           13.0%;  Score 651;  DB 1;  Length 1255;
  Best Local Similarity  27.1%;  Pred. No. 3.4e-19;
  Matches  290;  Conservative   79;  Mismatches  437;  Indels  266;  Gaps   52;

Qy      75 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 133
           | |||           | :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db      16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy     134 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTP----------KE 166
           |    :||:|    |    ||              | |     ||||           :
Db      66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVTSAPDNK 125

Qy     167 PTPTT-------IKSAPTTPKEPAPTTT-------KSAPTTPKEPAPTTTKEP------A 206
           | | :       : ||| |   ||| :|        ||| |   ||| :|  |      |
Db     126 PAPGSTAPPAHGVTSAPDT--RPAPGSTAPPAHGVTSAPDT--RPAPGSTAPPAHGVTSA 181

Qy     207 PTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTP----------KKPAPTTPKEP---- 252
           | |   ||| :|  ||   | ||| |   |  | |           :||| :   |
Db     182 PDT--RPAPGSTAPPAHGVT-SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGV 238

Qy     253 --APTTPKEPTPTTP----------KEPAPTTKEP-------APTTPKEPAPTAPK---- 289
             || |   | | |            ||| :  |       || |   |  |||
Db     239 TSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGV 298

Qy     290 KPAPTTPKEPAPTTPK----EPAPTTTKEPSPTTPKEPAPTT---TKSAPTTTKEPAPTT 342
           || |   | | |        || |   | | |      |:   |: || :|  ||
Db     299 TSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGV 358

Qy     343 TKSAPTTPKEPSPTTTKEP------APTTPKEPAPTTPK----KPAPTTPKEPAPTTPK- 391
           | ||| |   |:| :|  |       || |   | | |        || |   | | |
Db     359 T-SAPDT--RPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPA 415

Qy     392 ---EPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEE 448
              || |   |  |||     |:  :| |       |       || |   | | |
Db     416 HGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAP-- 473

Qy     449 LAPTTPEEPTPTTPEEPAPTTPKA----AAPNTPKEPAPTTPK----EPAPTTPKEPAPT 500
             |       | |   | | | |    :||:|   | | |        || |   | |
Db     474 --PAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGST 531

Qy     501 TPKETAPTTPKGTAPTTLKEPAPTTP----------KKPAPKELAPTTTKEPTSTTSDKP 550
           |     |     :|| |   |  | |           :|||     ||       ||   :|
Db     532 AP----PAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAP-PAHGVTSAPDTRP 586

Qy     551 AP--TTPKGTAPTTPKEPAPTTPKEPAPTTPKG----TAPTTLKEPAPTTP--------- 595
           ||  | |      |:   || |   |  | |        :|| |   |  | |
Db     587 APGSTAPPAHGVTS----APDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAP 642

Qy     596 -KKPAPKELAPTTTKGPTSTTSDKPAP--TTPK----ETAPTTPKEPAPTTP-------- 640
            :|||   ||    | ||   :|||  | |      :|| |   |  | |
Db     643 DTRPAPGSTAP-PAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSA 701

Qy     641 --KKPAPTTPETPPPTTSEVSTPTTTKEPTT-------IHKSPDESTPELSAEPTPKALE 691
             :||| :  | ||      | | |   | :       :  :||      |  |    :
Db     702 PDTRPAPGS--TAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT 759

Qy     692 NSPKEPGVPTTKTPAA--------TKPEMTTTAKDK---TTERDLRTTPETT-------T 733
           ::|     | :  | |         |:|   :||      |:  | |  | :|       |
Db     760 SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT 819

Qy     734 AAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTP--FKITTLKTTTLAPKVTTTKK 791
           :||    :|       |       |:   |       | |    :|:   |  ||  |
Db     820 SAP----DTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPA 875

Qy     792 TITTT--EIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKT 849
           |:  :    |  || |    |::  | | |   : ||      |  | | |    | :
Db     876 HGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAP--GSTAPPAHGVTSAPDTRP---APGS 930

Qy     850 TPTPRKMTSTMPELNP-----------------TSRIAEAMLQ--TTTRPNQTPNSK-- 887
           |  |    :: |:  |                  |  |  ::   |: |  || ||
Db     931 TAPPAHGVTSAPDNRPALGSTAPPVHNVTSASGSASGSASTLVHNGTSARATTTPASKST 990

Qy     888 -------------LVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDY 925
                         |   : |:: :       | :    |   |:::  : :
Db     991 PFSIPSHHSDTPTTLASHSTKTDASSTHHSSVPPLTSSNHSTSPQLSTGVSF 1042
```

```
RESULT   5
FP1_MYTED
ID   FP1_MYTED      STANDARD;      PRT;   875 AA.
AC   Q25460;
DT   30-MAY-2000 (Rel. 39, Created)
DT   30-MAY-2000 (Rel. 39, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ADHESIVE PLAQUE MATRIX PROTEIN (POLYPHENOLIC ADHESIVE PROTEIN) (FOOT
DE   PROTEIN 1) (MEFP1) (FRAGMENT).
GN   FP1.
OS   Mytilus edulis (Blue mussel).
OC   Eukaryota; Metazoa; Mollusca; Bivalvia; Pteriomorphia; Mytiloida;
OC   Mytiloidea; Mytilidae; Mytilus.
OX   NCBI_TaxID=6550;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=91025829; PubMed=1367451;
RA   Filpula D.R., Lee S.M., Link R.P., Strausberg S.L., Strausberg R.L.;
RT   "Structural and functional repetition in a marine mussel adhesive
```

```
          ||: :      ||  |:   || :  |     : :   |  | |||
Db   1167 VPTTYV------TTLAPSAPVTPATNAVPTTITTTECSAATNAAGET 1207


RESULT   4
MUC1_HUMAN
ID   MUC1_HUMAN     STANDARD;      PRT;  1255 AA.
AC   P15941; P15942; P13931; P17626; Q14128; Q16442; Q16437; Q9Y4J2;
DT   01-JAN-1990 (Rel. 13, Created)
DT   01-APR-1990 (Rel. 14, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   MUCIN 1 PRECURSOR (POLYMORPHIC EPITHELIAL MUCIN) (PEM) (PEMT)
DE   (EPISIALIN) (TUMOR-ASSOCIATED MUCIN) (CARCINOMA-ASSOCIATED MUCIN)
DE   (TUMOR-ASSOCIATED EPITHELIAL MEMBRANE ANTIGEN) (EMA) (H23AG) (PEANUT-
DE   REACTIVE URINARY MUCIN) (PUM) (BREAST CARCINOMA-ASSOCIATED ANTIGEN
DE   DF3).
GN   MUC1.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Pancreas;
RX   MEDLINE=90368716; PubMed=2394722;
RA   Lan M.S., Batra S.K., Qi W.-N., Metzgar R.S., Hollingsworth M.A.;
RT   "Cloning and sequencing of a human pancreatic tumor mucin cDNA.";
RL   J. Biol. Chem. 265:15294-15299(1990).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=90202794; PubMed=2318825;
RA   Ligtenberg M.J.L., Vos H.L., Gennissen A.M.C., Hilkens J.;
RT   "Episialin, a carcinoma-associated mucin, is generated by a
RT   polymorphic gene encoding splice variants with alternative amino
RT   termini.";
RL   J. Biol. Chem. 265:5573-5578(1990).
RN   [3]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Breast carcinoma;
RX   MEDLINE=90368715; PubMed=1697589;
RA   Gendler S.J., Lancaster C.A., Taylor-Papadimitriou J., Duhig T.,
RA   Peat N., Burchell J., Pemberton L., Lalani E.-N., Wilson D.;
RT   "Molecular cloning and expression of human tumor-associated
RT   polymorphic epithelial mucin.";
RL   J. Biol. Chem. 265:15286-15293(1990).
RN   [4]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=91097524; PubMed=2268309;
RA   Lancaster C.A., Peat N., Duhig T., Wilson D.,
RA   Taylor-Papadimitriou J., Gendler S.J.;
RT   "Structure and expression of the human polymorphic epithelial mucin
RT   gene: an expressed VNTR unit.";
RL   Biochem. Biophys. Res. Commun. 173:1019-1029(1990).
RN   [5]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Breast carcinoma;
RX   MEDLINE=90276413; PubMed=2351132;
RA   Wreschener D.H., Hareuveni M., Tsarfaty I., Smorodinsky N.,
RA   Horov J., Zaretsky J., Kotkes P., Weiss M., Lathe R., Dion A.,
RA   Keydar I.;
RT   "Human epithelial tumor antigen cDNA sequences. Differential splicing
RT   may generate multiple protein forms.";
RL   Eur. J. Biochem. 189:463-473(1990).
RN   [6]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Breast carcinoma;
RX   MEDLINE=90276414; PubMed=2112460;
RA   Hareuveni M., Tsarfaty I., Zaretsky J., Kotkes P., Horev J.,
RA   Zrihan S., Weiss M., Green S., Lathe R., Keydar I., Wreschner D.H.;
RT   "A transcribed gene, containing a variable number of tandem repeats,
RT   codes for a human epithelial tumor antigen. cDNA cloning, expression
RT   of the transfected gene and over-expression in breast cancer
RT   tissue.";
RL   Eur. J. Biochem. 189:475-486(1990).
RN   [7]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=91033045; PubMed=1688329;
RA   Tsarfaty I., Hareuveni M., Horev J., Zaretsky J., Weiss M.,
RA   Jeltsch J.M., Garnier J.M., Lathe R., Keydar I., Wreschner D.H.;
RT   "Isolation and characterization of an expressed hypervariable gene
RT   coding for a breast-cancer-associated antigen.";
RL   Gene 93:313-318(1990).
RN   [8]
RP   PARTIAL SEQUENCE FROM N.A.
RX   MEDLINE=88330762; PubMed=3417635;
RA   Gendler S.J., Taylor-Papadimitriou J., Duhig T., Rothbard J.,
RA   Burchell J.;
RT   "A highly immunogenic region of a human polymorphic epithelial mucin
RT   expressed by carcinomas is made up of tandem repeats.";
RL   J. Biol. Chem. 263:12820-12823(1988).
RN   [9]
RP   SEQUENCE OF 1-169 FROM N.A.
RX   MEDLINE=90088473; PubMed=2597151;
RA   Abe M., Siddiqui J., Kufe D.;
RT   "Sequence analysis of the 5' region of the human DF3 breast
RT   carcinoma-associated antigen gene.";
RL   Biochem. Biophys. Res. Commun. 165:644-649(1989).
RN   [10]
RP   SEQUENCE OF 1-109 FROM N.A.
RC   TISSUE=Thyroid;
RX   MEDLINE=96183746; PubMed=8608966;
RA   Weiss M., Baruch A., Keydar I., Wreschner D.H.;
RT   "Preoperative diagnosis of thyroid papillary carcinoma by reverse
RT   transcriptase polymerase chain reaction of the MUC1 gene.";
RL   Int. J. Cancer 66:55-59(1996).
RN   [11]
RP   SEQUENCE OF 1-89 FROM N.A.
RC   TISSUE=Lung;
RX   MEDLINE=96181716; PubMed=8604237;
RA   Yu C.J., Yang P.C., Shew J.Y., Hong T.M., Yang S.C., Lee Y.C.,
RA   Lee L.N., Luh K.T., Wu C.W.;
RT   "Mucin mRNA expression in lung adenocarcinoma cell lines and
RT   tissues.";
RL   Oncology 53:118-126(1996).
RN   [12]
RP   SEQUENCE OF 1-46 FROM N.A.
RC   TISSUE=Breast carcinoma;
RA   Buluwela L., Liu Q., Luqmani Y.A., Gomm J.J., Coombes R.C.;
RL   Submitted (OCT-1992) to the EMBL/GenBank/DDBJ databases.
CC   -!- FUNCTION: DIRECT OR INDIRECT INTERACTION WITH ACTIN
CC       CYTOSKELETON.
CC   -!- SUBCELLULAR LOCATION: TYPE I MEMBRANE PROTEIN. A SECRETED FORM
CC       IS ALSO PRODUCED.
CC   -!- ALTERNATIVE PRODUCTS: VARIOUS VARIANTS ARE PRODUCED BY ALTERNATIVE
CC       SPLICING.
CC   -!- TISSUE SPECIFICITY: ABERRANTLY EXPRESSED IN HUMAN EPITHELIAL
CC       TUMORS, SUCH AS BREAST CANCER.
CC   -!- PTM: HIGHLY GLYCOSYLATED (N-AND O-LINKED CARBOHYDRATES AND SIALIC
CC       ACID).
CC   -!- POLYMORPHISM: THE NUMBER OF REPEATS IS HIGHLY POLYMORPHIC. IT
CC       VARIES FROM 21 TO 125 IN THE NORTHERN EUROPEAN POPULATION. THE
CC       MOST FREQUENT ALLELES CONTAINS 41 AND 85 REPEATS.
CC   -!- SIMILARITY: CONTAINS 1 SEA DOMAIN.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; J05582; AAA60019.1; -.
DR   EMBL; M32738; AAA35804.1; -.
DR   EMBL; M32739; AAA35806.1; -.
DR   EMBL; J05581; AAA59876.1; -.
```

```
Qy    871 EAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLP 927
                 :|  ||    :|  |   |  || |      || :       ||  : |
Db   1425 --------KPTSTPAPTEIE-EPTPSDVPGAIGGEHRAYLRGY-------PDGSFRP 1465


RESULT   3
AMYH_YEAST
ID   AMYH_YEAST     STANDARD;      PRT;  1367 AA.
AC   P08640; P08068;
DT   01-AUG-1988 (Rel. 08, Created)
DT   01-FEB-1995 (Rel. 31, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   GLUCOAMYLASE S1/S2 PRECURSOR (EC 3.2.1.3) (GLUCAN 1,4-ALPHA-
DE   GLUCOSIDASE) (1,4-ALPHA-D-GLUCAN GLUCOHYDROLASE).
GN   STA1 OR STA2 OR MAL5 OR YIR019C.
OS   Saccharomyces cerevisiae (Baker's yeast).
OC   Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
OC   Saccharomycetales; Saccharomycetaceae; Saccharomyces.
OX   NCBI_TaxID=4932;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=S288C / AB972;
RA   Barrell B.G., Badcock K., Bankier A.T., Bowman S., Brown D.,
RA   Churcher C.M., Connor R., Copsey T., Dear S., Devlin K., Fraser A.,
RA   Gentles S., Hamlyn N., Horsnell T.S., Hunt S., Jagels K., Jones M.,
RA   Louis E., Lye G., Moule S., Moule T., Odell C., Pearson D.,
RA   Rajandream M.A., Riles L., Rowley N., Skelton J., Smith V.,
RA   Walsh S.V., Whitehead S.;
RL   Submitted (DEC-1994) to the EMBL/GenBank/DDBJ databases.
RN   [2]
RP   SEQUENCE OF 1-242 AND 762-1331 FROM N.A.
RX   MEDLINE=87194600; PubMed=3106330;
RA   Yamashita I., Nakamura M., Fukui S.;
RT   "Gene fusion is a possible mechanism underlying the evolution of
RT   STA1.";
RL   J. Bacteriol. 169:2142-2149(1987).
RN   [3]
RP   SEQUENCE OF 1-31 FROM N.A.
RC   STRAIN=SPX101-1C;
RX   MEDLINE=89031230; PubMed=3141213;
RA   Pardo J.M., Ianez E., Zalacain M., Claros M.G., Jimenez A.;
RT   "Similar short elements in the 5' regions of the STA2 and SGA genes
RT   from Saccharomyces cerevisiae.";
RL   FEBS Lett. 239:179-184(1988).
CC   -!- CATALYTIC ACTIVITY: HYDROLYSIS OF TERMINAL 1,4-LINKED ALPHA-D-
CC       GLUCOSE RESIDUES SUCCESSIVELY FROM NON-REDUCING ENDS OF THE CHAINS
CC       WITH RELEASE OF BETA-D-GLUCOSE.
CC   -!- SIMILARITY: TO S.POMBE SPBC215.13.
CC   -!- SIMILARITY: SOME, TO S.POMBE SPCC285.13C.
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; Z38061; CAA86176.1; -.
DR   EMBL; M16164; AAA35014.1; -.
DR   EMBL; M16165; AAA35015.1; -.
DR   EMBL; X13857; CAA32069.1; -.
DR   PIR; B26877; B26877.
DR   PIR; A26877; A26877.
DR   PIR; S48478; S48478.
DR   SGD; S0001458; MUC1.
KW   Hydrolase; Glycosidase; Polysaccharide degradation; Glycoprotein;
KW   Signal; Multigene family.
FT   SIGNAL        1     21       POTENTIAL.
FT   CHAIN        22   1367       GLUCOAMYLASE S1/S2.
FT   DOMAIN      210   1367       SER/THR-RICH.
FT   CARBOHYD    817    817       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    874    874       N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE   1367 AA;  136110 MW;  91C00E2DBD61AA9D CRC64;


  Query Match           15.2%;  Score 762.5;  DB 1;  Length 1367;
  Best Local Similarity  27.8%;  Pred. No. 1.7e-23;
  Matches  280;  Conservative  115;  Mismatches  441;  Indels  171;  Gaps   42;

Qy     35 TTPDTSTTQHNKVSTSPKITTAKPINP--------RPSLPPNSDTSKETSLTVNKETTVE 86
          :| ::|||  :   :|   :|  | |          :|: |  :  :||     : :||
Db    236 STSESSTTTSSTSESSTSSSTTAPATPTTTSCTKEKPTPPTTTSCTKEKPTPPHHDTTPC 295

Qy     87 TKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETT--TKGPA 144
          ||: |||:| | |   :|||:   | |   |:    ||       ||| : ||  :  |
Db    296 TKKKTTTSK-TCT---KKTTTPVPTPS-SSTTESSSAPV------PTPSSSTTESSSAPV 344

Qy    145 LTTPKE------PTPTTPKEPASTTPKEPTPTTIKSAP----TTPKEPAPTTTKSAPTTP 194
          ::  |      |||::    :|: |   : |  |||    ||    ||  | |: ||
Db    345 TSSTTESSSAPVPTPSSSTTESSSAPVTSSTTESSSAPVTSSTTESSSAPVPTPSSSTTE 404

Qy    195 KEPAP----TTTKEPAP----TTPKEPAP----TTTKEPAP----TTTKSAPTTPKEPAP 238
             ||    ||    ||    ||    ||    ||    ||    ||  |:   |   :
Db    405 SSSAPVTSSTTESSSAPVTSSTTESSSAPVTSSTTESSSAPVTSSTTESSSAPVPTPSSS 464

Qy    239 TTPKKPAPTTPKEPAPTTPKEPTP---TTPKEPAPTTKE-------PAPT----TPKEPA 284
          ||    || |       ::   |||   ||    || |           | ||    | :  :
Db    465 TTESSSAPVTSSTTESSSAPVPTPSSSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSS 524

Qy    285 PTAPKKPAPTTPKEPAPTTPKE--------PAPTTTKEPSPTTPKEPAPTTTKSAPTTTK 336
             ||   : ||    || |             | |:::   | :||   : | : |||  |
Db    525 APAPTPSSSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSSTPVTSSTTESSSAPVPT- 583

Qy    337 EPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTT------PKEPAPTTP 390
           |: :||:|:      || :||:  :   |   : ||    || |          | ||
Db    584 -PSSSTTESSSAPVPTPSSSTTESSSAPAPTPSSSTTESSSAPVTSSTTESSSAPVPT-- 640

Qy    391 KEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAP----TTPEKPAPTTP 446
           |: :||:  :   |   : ||   :||  |   : ||    ||    ||    || |
Db    641 --PSSSTTESSSAPVPTPSSSTTESSSAP-VPTPSSSTTESSSAPVTSSTTESSSAPVTS 697

Qy    447 EELAPTTPEEPTP---TTPEEPAPT-TPKAAAPNTPKEPAP-----TTPKEPAPTT---- 493
             ::   |||   ||    ||  || ::   :   | |     ||    || |
Db    698 STTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESSSAPVTSSTT 757

Qy    494 -------PKEPAPTTPKETAPT-TPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTST 545
                |   : ||   :||  || :   :   | ||        :|| |   :|
Db    758 ESSSAPVPTPSSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESSVAPVPTPSSSSN 817

Qy    546 TSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAP 605
           :     :||  ::  :   | ||     : ||   :|| :       :   | |  :
Db    818 ITSSAPSSTPFSSSTESSSVPVPT---PSSSTTESSSAPVSSSTTESSVAPVPTPSSSSN 874

Qy    606 TTTKGP-----TSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPET--PPPTTSEV 658
           |:  |     :|||       ||      |:: | |   |    : ||  |  |  ||: |
Db    875 ITSSAPSSIPFSSTTESFSTGTT---VTPSSSKYPGSQTETSVSSTTETTIVPTKTTTSV 931

Qy    659 STPTTTKEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTT---- 714
          :||:||   ||:  :   |  | ::  :|| :  :      ||||  | : |    ||
Db    932 TTPSTTTITTTVCSTGTNSAGETTSGCSPKTVTTT-----VPTTTTTSVTTSSTTTITTT 986

Qy    715 ---TAKDKTTERDLRTTPET-TTAAPKMTKETATTTEKTTESKITATTTQVTST--TTQD 768
             |  :   |     :|:| ||  |  |  :  | :| || |   |  || |::|  ||:
Db    987 VCSTGTNSAGETTSGCSPKTITTTVPCSTSPSETASESTTTSPTTPVTTVVSTTVVTTEY 1046

Qy    769 TTPFK------------ITTLKTTTLA--PKVTT----TKKTITTTEI---MNKPEET- 805
          :| |               | |   ||:|  | |||     |  |||||       |   ||
Db   1047 STSTKPGGEITTTFVTKNIPTTYLTTIAPTPSVTTVTNFTPTTITTTVCSTGTNSAGETT 1106

Qy    806 --AKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPEL 863
             ||  |   :|  :  |:|    |:      :       |   :  |   :
Db   1107 SGCSPKTVTTTVPCSTGTGEYTTEATTLVTTAVTTTVVTTESSTGTNSAGKTTTGYTTKS 1166

Qy    864 NPTSRIAEAMLQTTTRPNQ--TPNSKLVEVNPKSED---AGGAEGET 905
```

```
Qy     492 TTPKEPAPT-TPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKP 550
           ||   | || |  :|  |||  |  |    | ||  : |    :  |||  || | :
Db    1950 TTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTTVTPTPTPTGTQ 2009

Qy     551 AP-TTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTK 609
            | |||  |  |    | ||  : |  |||  |  |    | ||  : |    :  |||
Db    2010 TPTTTPITTTTTVTPTPTPTGTQTPT-TTPITTTTTVTPTPTPTGTQTPTTTPITTTTTV 2068

Qy     610 GPTSTTSDKPAP-TTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPT 668
           || | :    | |||  |  |    | ||  : |  ||| |   | :   ||| |: ||
Db    2069 TPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPT-TTPITTTTTVTPTPTPTGTQTPT 2127

Qy     669 TIHKSPDESTPELSAEPTPKALENSPKEP----------GVPT-TKTPAATKPEMTTTAK 717
           |   :|  :|  ::  |||   :    |              || |:|| |    |||
Db    2128 T---TPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVT 2184

Qy     718 DKTTERDLR---TTP--ETTTAAPKMT---KETATTTEKTTESKITATTTQVTSTTTQDT 769
           |    :   |||   |||  |  |    :| |||  || : :| | |  | | |  |
Db    2185 PTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTP-TGTQTPTT 2243

Qy     770 TPFKITTLKTTTLAPKVTTT-KKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKA 828
           ||  |||  |||: |  | |  :| ||| |      |  |    | :  |||     |
Db    2244 TP--ITT--TTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVT 2299

Qy     829 P-KKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNP---TSRIAEAMLQTTTR-PNQT 883
           |  || |: | | |       ||||    :  |   |: :      | |: |  |
Db    2300 PTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTT 2359

Qy     884 PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTP 921
           | :    | |     |     :||     |      |||
Db    2360 PITTTTTVTPTPTPTG---TQTP--TTTPITTTTTVTP 2392


RESULT   2
SLP1_CLOTM
ID   SLP1_CLOTM     STANDARD;      PRT;  1664 AA.
AC   Q06852;
DT   01-JUN-1994 (Rel. 29, Created)
DT   01-JUN-1994 (Rel. 29, Last sequence update)
DT   01-FEB-1996 (Rel. 33, Last annotation update)
DE   CELL SURFACE GLYCOPROTEIN 1 PRECURSOR (OUTER LAYER PROTEIN B) (S-LAYER
DE   PROTEIN 1).
GN   OLPB.
OS   Clostridium thermocellum.
OC   Bacteria; Firmicutes; Bacillus/Clostridium group; Clostridiaceae;
OC   Clostridium.
OX   NCBI_TaxID=1515;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=NCIB 10682;
RX   MEDLINE=93209931; PubMed=8458832;
RA   Fujino T., Beguin P., Aubert J.-P.;
RT   "Organization of a Clostridium thermocellum gene cluster encoding the
RT   cellulosomal scaffolding protein CipA and a protein possibly involved
RT   in attachment of the cellulosome to the cell surface.";
RL   J. Bacteriol. 175:1891-1899(1993).
CC   -!- SUBUNIT: ASSEMBLED INTO MONO-LAYERED CRYSTALLINE ARRAYS.
CC   -!- SUBCELLULAR LOCATION: CELL WALL.
CC   -!- SIMILARITY: CONTAINS 4 S-LAYER HOMOLOGY (SLH) DOMAINS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X67506; CAA47841.1; -.
DR   InterPro; IPR001119; SLH.
DR   Pfam; PF00395; SLH; 3.
DR   PROSITE; PS01072; SLH_DOMAIN; 2.
KW   Cell wall; S-layer; Signal; Repeat.
FT   SIGNAL       1     28       POTENTIAL.
FT   CHAIN       29   1664       CELL SURFACE GLYCOPROTEIN 1.
FT   DOMAIN      36    763       4 X 156 AA APPROXIMATE REPEATS.
FT   REPEAT      36    191       1.
FT   REPEAT     207    363       2.
FT   REPEAT     409    565       3.
FT   REPEAT     607    763       4.
FT   DOMAIN     771   1377       APPROXIMATE TANDEM REPEATS OF
FT                               T-P-S-D-E-P.
FT   DOMAIN    1378   1449       GLY/PRO/SER/THR-RICH.
FT   DOMAIN    1453   1494       SLH 1 (INCOMPLETE).
FT   DOMAIN    1495   1565       SLH 2.
FT   DOMAIN    1566   1625       SLH 3.
FT   DOMAIN    1626   1646       SLH 4 (INCOMPLETE).
SQ   SEQUENCE   1664 AA;  178194 MW;  5F396695BA9FE74B CRC64;


  Query Match             18.8%;  Score 942;  DB 1;  Length 1664;
  Best Local Similarity   33.0%;  Pred. No. 2.2e-30;
  Matches  276;  Conservative   83;  Mismatches  312;  Indels  166;  Gaps   43;

Qy     128 VLAKPTP-KAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTT 186
           |: :| | || :    |  |   ||||:           ||||:     ||   || |:
Db     758 VVIQPAPIKAASDEPIPTDTPSDEPTPS----------DEPTPS---DEPTPSDEPTPSD 804

Qy     187 TKSAPTTPKEPAPTTTKEPAPTTPKEPA----PTTTKEPAPTTTKSAPTTPKEPAPTTPK 242
             :  ||:|| || |    ||   ||     || : || |:   :   ||:|| ||
Db     805 EPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTP 864

Qy     243 KPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPT-TPKEPAPTAPKKPAPTTPKEPAP 301
           ||  || |:   ||||:   || | : || |: ||:|| ||       ||   || |
Db     865 SDEPTPSDEPTPS--DEPTPS--DEPTP-SDEPTPSETPEEPIPTDTPSDEPTPSDEPTP 919

Qy     302 TTPKEPA--PTTTKEPSPT-TPKEPAPTTTKS-APTTTKEPAPTTTKSAPTTPKEPSPTT 357
           :   |:  || : ||:|: ||:|| || | |  || : || |:           ||
Db     920 SDEPTPSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPS----------DEPTP 968

Qy     358 TKEPAPTTPKEPAPTTPKKPAPTTPKEPAPT-TPKEPAPTTTKKPAPTAPKEPAPTTPKE 416
           : || |:   || |:      ||   || |: ||:|| || |    ||   || |:
Db     969 SDEPTPS--DEPTPS----DEPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPSD--- 1019

Qy     417 TAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPT-TPEEPAPTTPKAAAP 475
                   ||  ::  ||  ::| |:      ||  :|||| |||| ||   :  |
Db    1020 ---------EPTPSDE--PTPSDEPTPSD----EPTPSDEPTPSETPEEPIPTDTPSDEP 1064

Qy     476 NTPKEPAPTTPKEPA--PTTPKEPAPT-TPKETAPTTPKGTAPTTLKEPAPTTPKKPAPK 532
              || |:    |:  ||   || |: ||:|  ||      ||   || |:   :| |
Db    1065 TPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPS--DEPTPS 1122

Qy     533 ELAPTTTKEPT-STTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPA 591
           :  || : ||| | | ::| ||      ||   || |:   || |:      ||   ||
Db    1123 D-EPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPS--DEPTPS----DEPTPSDEPT 1175

Qy     592 PT-TPKKPAPKELAPTTTKGPTSTTSDKPAPT---TPKETAPTTPKEPAPT-TPKKPAPT 646
           |: ||::|      || |     | ||:| |:   || :  ||   || |: ||::| ||
Db    1176 PSETPEEP-----IPTDTPSDEPTPSDEPTPSDEPTPSD-EPTPSDEPTPSETPEEPIPT 1229

Qy     647 TPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPA 706
             : || |:   || : |||       || ||  | ||||     :|:|| :||
Db    1230 DTPSDEPTPSD--EPTPSDEPT----PSDEPTP--SDEPTP---SETPEEP-IPTDTPSD 1277

Qy     707 ATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTT 766
           |     |  |: |  |  |  :  | : : |          | :| |  |  |     | :
Db    1278 EPTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPS 1337

Qy     767 QDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ-KP 825
           : ||    |                        | :   | |  | |   ||  :  |
Db    1338 DEPTPSDEPT----------------------PSDEPTPSDEPTPSDEPTPSDEPTP 1372

Qy     826 TKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTS--------------TMPELNPTSRIA 870
           :: |::|| |  |         | ||||     |               | |   |||
Db    1373 SETPEEPTPTTTP-----TPTPSTTPTSGSGGSGGSGGGGGGGGGTVPTSPTPTPTS--- 1424
```

```
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   -----------------------------------------------------------------------
DR   EMBL; L21998; AAB95295.1; -.
DR   EMBL; M74027; AAA59875.1; -.
DR   EMBL; M94131; AAA59163.1; -.
DR   EMBL; M94132; AAA59164.1; -.
DR   MIM; 158370; -.
DR   InterPro; IPR000359; Cys_knot.
DR   InterPro; IPR000561; EGF-like.
DR   InterPro; IPR002400; GF_cysknot.
DR   InterPro; IPR001007; VWFC.
DR   InterPro; IPR001846; Vwd.
DR   Pfam; PF00007; Cys_knot; 1.
DR   Pfam; PF00094; vwd; 4.
DR   PRINTS; PR00438; GFCYSKNOT.
DR   SMART; SM00214; VWC; 2.
DR   SMART; SM00011; VWC_def; 2.
DR   SMART; SM00216; VWD; 4.
DR   PROSITE; PS00022; EGF_1; UNKNOWN_1.
DR   PROSITE; PS01185; CTCK_1; 1.
DR   PROSITE; PS01225; CTCK_2; 1.
DR   PROSITE; PS01208; VWFC; 2.
KW   Glycoprotein; Repeat; Signal.
FT   SIGNAL        1     20       POTENTIAL.
FT   CHAIN        21   5179       MUCIN 2.
FT   DOMAIN     1401   1747       APPROXIMATE REPEATS.
FT   REPEAT     1401   1416       1.
FT   REPEAT     1417   1432       2.
FT   REPEAT     1433   1448       3.
FT   REPEAT     1449   1464       4.
FT   REPEAT     1465   1471       5.
FT   REPEAT     1472   1478       6.
FT   REPEAT     1479   1494       7A.
FT   REPEAT     1495   1517       7B.
FT   REPEAT     1518   1533       8A.
FT   REPEAT     1534   1556       8B.
FT   REPEAT     1557   1572       9A.
FT   REPEAT     1573   1596       9B.
FT   REPEAT     1597   1612       10A.
FT   REPEAT     1613   1635       10B.
FT   REPEAT     1636   1651       11A.
FT   REPEAT     1652   1675       11B.
FT   REPEAT     1676   1683       12.
FT   REPEAT     1684   1699       13.
FT   REPEAT     1700   1715       14.
FT   REPEAT     1716   1731       15.
FT   REPEAT     1732   1747       16.
FT   DOMAIN     4815   4886       VWFC 1.
FT   DOMAIN     4924   4991       VWFC 2.
FT   DOMAIN     5075   5160       CTCK.
FT   DISULFID   5075   5122       BY SIMILARITY.
FT   DISULFID   5089   5136       BY SIMILARITY.
FT   DISULFID   5098   5152       BY SIMILARITY.
FT   DISULFID   5102   5154       BY SIMILARITY.
FT   DISULFID      ?   5159       BY SIMILARITY.
FT   CARBOHYD    163    163       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    423    423       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    670    670       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    770    770       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    894    894       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1139   1139       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1154   1154       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1215   1215       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1230   1230       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1246   1246       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1787   1787       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1820   1820       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4339   4339       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4351   4351       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4362   4362       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4373   4373       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4422   4422       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4438   4438       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4502   4502       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4616   4616       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4627   4627       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4752   4752       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4787   4787       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4881   4881       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4888   4888       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4955   4955       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4970   4970       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   5019   5019       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   5038   5038       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   5069   5069       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CONFLICT   1351   1351       H -> L (IN REF. 3).
FT   CONFLICT   1412   1412       T -> S (IN REF. 3).
FT   CONFLICT   1449   1449       L -> P (IN REF. 3).
FT   CONFLICT   1504   1504       M -> T (IN REF. 3).
FT   CONFLICT   4192   4192       G -> S (IN REF. 2).
SQ   SEQUENCE   5179 AA;  540295 MW;  85CD7571FB9A5663 CRC64;
```

```
  Query Match             23.3%;  Score 1168;  DB 1;  Length 5179;
  Best Local Similarity   32.5%;  Pred. No. 9.1e-39;
  Matches  363;  Conservative   53;  Mismatches  448;  Indels  254;  Gaps   40;

Qy      21 DEAGSGLDNGDFK----------------VTTPDTSTTQH-NKVSTSPK---------- 52
           |   || |:|| :                |  |  |  ||  ||
Db    1312 DHPSSGSDDGDREPFDGVCGAPEDIECRSVKDPHLSLEQHGQKVQCDVSVGFICKNEDQF 1371

Qy      53 ------------------------ITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETK 88
                                    |||  |   || || : |:   : | :  ||
Db    1372 GNGPFGLCYDYKIRVNCCWPMDKCITTPSPPTTTPSPPPTTTTTLPPTTTPSPPTT---- 1427

Qy      89 ETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTP 148
            ||||   |:|     ||:                 ||:     |:|   |||  |  |||
Db    1428 -TTTTPPPTTTPSPPITTTTTP-----------LPTT----TPSPPISTTTTPPPTTTP 1470

Qy     149 KEPT-----PTTPKEPASTTPKEPTPTTIKSAP-TTPKEP-------APTTTKSAP-TTP 194
             ||      |||   | :||   | |||  | | |||  |        ||||  | |  ||
Db    1471 SPPTTTPSPPTTTPSPPTTTTTTPPPTTTPSPPMTTPITPPASTTTLPPTTTPSPPTTTT 1530

Qy     195 KEPAPTTTKEPAPTTPKEP-------APTTTKEPAPTTTKSAP--TTPKEPAPTTPKKPA 245
             | ||||  |  |||  |         ||||  | |||| : |  |||  |  |||  |
Db    1531 TTPPPTTTPSPPTTTPITPPTSTTTLPPTTTPSPPPTTTTPPPTTTPSPPTTTTPSPPT 1590

Qy     246 PTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEP-------APTAPKKPAPTTPKE 298
            ||   |  |||  || ||   |  ||  |  |||  |          ||    | |||
Db    1591 ITTTTPPPTTTPSPPTTTTTPPPTTTPSPPTTTPITPPTSTTTLPPTTTPSPPPTTTTT 1650

Qy     299 PAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAP---TTPKEPSP 355
           | |||   |  |||  |  ||   | |||| |:| ||    |||| : |   |||  |
Db    1651 PPPTTTPSPPTTTTPSPPITTTTTPPPTTTPSSPITTTPSPPTTTMTTPSPTTTPSSPI- 1709

Qy     356 TTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTP- 414
           |||  |: ||   | |||   |:|||   |  ||     ||||  |  | |  |: | |
Db    1710 TTTTTPSSTTTPSPPPTTMTTPSPTTTPSPPTTTMTTLPPTTTSSPLTTTPLPPSITPPT 1769

Qy     415 ----KETAPTTP-------------------------------------------KKLTP 427
               | ||||                                             : |
Db    1770 FSPFSTTTPTTPCVPLCNWTGWLDSGKPNFHKPGGDTELIGDVCGPGWAANISCRATMYP 1829

Qy     428 ------------------------------------------------------TTPEKLAPTT 437
                                                                 | |  :  ||
Db    1830 DVPIGQLGQTVVCDVSVGLICKNEDQKPGGVIPMAFCLNYEINVQCCECVTQPTTMTTTT 1889

Qy     438 PEKPAPTTPEELAPTTPEEPTPT-TPEEPAPTTPKAAAPNTPKEPAPTTPKEP-----AP 491
           | | | |   :  ||   |||| |  :   |||          | ||  : |
Db    1890 TENPTPPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITT 1949
```

GenCore version 4.5
Copyright (c) 1993 - 2000 Compugen Ltd.

OM protein - protein search, using sw model

Run on: April 26, 2002, 16:13:35 ; Search time 47.4 Seconds
(without alignments)
727.882 Million cell updates/sec

Title: US-09-556-246-1_COPY_200_1140
Perfect score: 5011
Sequence: 1 VKDNKKNRTKKKPTPKPPVV.....DMDYLPRVPNQGIIINPMLS 941

Scoring table: BLOSUM62
Gapop 10.0 , Gapext 0.5

Searched: 100059 seqs, 36664827 residues

Total number of hits satisfying chosen parameters: 100059

Minimum DB seq length: 0
Maximum DB seq length: 2000000000

Post-processing: Minimum Match 0%
Maximum Match 100%
Listing first 45 summaries

Database : SwissProt_39:*

Pred. No. is the number of results predicted by chance to have a score greater than or equal to the score of the result being printed, and is derived by analysis of the total score distribution.

SUMMARIES

Result No.	Score	% Query Match	Length	DB	ID	Description
1	1168	23.3	5179	1	MUC2_HUMAN	Q02817 homo sapien
2	942	18.8	1664	1	SLP1_CLOTM	Q06852 clostridium
3	762.5	15.2	1367	1	AMYH_YEAST	P08640 saccharomyc
4	651	13.0	1255	1	MUC1_HUMAN	P15941 h mucin 1 p
5	556.5	11.1	875	1	FP1_MYTED	Q25460 mytilus edu
6	555.5	11.1	2700	1	ZAN_HUMAN	Q9y493 homo sapien
7	551	11.0	620	1	EXTN_TOBAC	P13983 nicotiana t
8	533	10.6	1087	1	NFH_MOUSE	P19246 mus musculu
9	530.5	10.6	1162	1	TCNA_TRYCR	P23253 trypanosoma
10	530	10.6	865	1	CPN_DROME	Q02910 drosophila
11	518.5	10.3	872	1	FP1_MYTCO	Q25434 mytilus cor
12	509.5	10.2	662	1	MUC1_XENLA	Q05049 xenopus lae
13	503.5	10.0	1970	1	RPB1_HUMAN	P24928 homo sapien
14	498.5	9.9	1970	1	RPB1_MOUSE	P08775 mus musculu
15	497	9.9	831	1	NFH_RAT	P16884 rattus norv
16	493.5	9.8	467	1	RPB1_CRIGR	P11414 cricetulus
17	488.5	9.7	826	1	SSP2_PLAYO	Q01443 plasmodium
18	488	9.7	1020	1	NFH_HUMAN	P12036 homo sapien
19	475.5	9.5	267	1	EXTN_MAIZE	P14918 zea mays (m
20	468.5	9.3	634	1	HWP1_CANAL	P46593 candida alb
21	467.5	9.3	5376	1	ZAN_MOUSE	O88799 mus musculu
22	467	9.3	817	1	VRP1_YEAST	P37370 saccharomyc
23	454	9.1	1161	1	YJ9P_YEAST	P47179 saccharomyc
24	452.5	9.0	2142	1	BAT2_HUMAN	P48634 homo sapien
25	448.5	9.0	670	1	VG50_HSV11	Q00130 ictalurid h
26	448	8.9	797	1	VGLX_HSVEB	P28968 equine herp
27	442.5	8.8	751	1	FP1_MYTGA	Q27409 mytilus gal
28	439.5	8.8	1083	1	T2D3_HUMAN	O00268 homo sapien
29	432.5	8.6	439	1	XP2_XENLA	P17437 xenopus lae
30	432	8.6	3164	1	TEGU_HSV11	P10220 herpes simp
31	426	8.5	2715	1	TRX2_HUMAN	Q9umn6 homo sapien
32	424.5	8.5	1125	1	MAP4_MOUSE	P27546 mus musculu
33	424	8.5	307	1	SGS3_DROME	P02840 drosophila
34	424	8.5	1251	1	YQU3_CAEEL	Q09550 caenorhabdi
35	422.5	8.4	1794	1	YAV1_SCHPO	Q10172 schizosacch
36	422.5	8.4	3421	1	TEGU_HSVEB	P28955 equine herp
37	419.5	8.4	2476	1	ZAN_PIG	Q28983 sus scrofa
38	419	8.4	1185	1	DRPL_HUMAN	P54259 homo sapien
39	415	8.3	1229	1	N121_HUMAN	Q9y2n3 homo sapien
40	411.5	8.2	907	1	VGP3_EBV	P03200 epstein-bar
41	410	8.2	1183	1	DRPL_RAT	P54258 rattus norv
42	406	8.1	2774	1	MAPA_RAT	P34926 rattus norv
43	406	8.1	3256	1	KI67_HUMAN	P46013 homo sapien
44	405	8.1	1411	1	TCOF_HUMAN	Q13428 homo sapien
45	400.5	8.0	2517	1	NCR2_HUMAN	Q9y618 h nuclear r

ALIGNMENTS

RESULT 1
MUC2_HUMAN
ID MUC2_HUMAN STANDARD; PRT; 5179 AA.
AC Q02817; Q14878;
DT 01-JUN-1994 (Rel. 29, Created)
DT 01-NOV-1997 (Rel. 35, Last sequence update)
DT 20-AUG-2001 (Rel. 40, Last annotation update)
DE MUCIN 2 PRECURSOR (INTESTINAL MUCIN 2).
GN MUC2 OR SMUC.
OS Homo sapiens (Human).
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX NCBI_TaxID=9606;
RN [1]
RP SEQUENCE FROM N.A.
RC TISSUE=Intestine;
RX MEDLINE=94132002; PubMed=8300571;
RA Gum J.R. Jr., Hicks J.W., Toribara N.W., Siddiki B., Kim Y.S.;
RT "Molecular cloning of human intestinal mucin (MUC2) cDNA.
RT Identification of the amino terminus and overall sequence similarity
RT to prepro-von Willebrand factor.";
RL J. Biol. Chem. 269:2440-2446(1994).
RN [2]
RP SEQUENCE OF 626-1895 AND 4196-5179 FROM N.A.
RC TISSUE=Colon;
RX MEDLINE=93016075; PubMed=1400449;
RA Gum J.R. Jr., Hicks J.W., Toribara N.W., Rothe E.-M., Lagace R.E.,
RA Kim Y.S.;
RT "The human MUC2 intestinal mucin has cysteine-rich subdomains located
RT both upstream and downstream of its central repetitive region.";
RL J. Biol. Chem. 267:21375-21383(1992).
RN [3]
RP SEQUENCE OF 1343-1895 AND 4176-4195 FROM N.A.
RX MEDLINE=91358717; PubMed=1885763;
RA Toribara N.W., Gum J.R. Jr., Culhane P.J., Lagace R.E., Hicks J.W.,
RA Petersen G.M., Kim Y.S.;
RT "MUC-2 human small intestinal mucin gene structure. Repeated arrays
RT and polymorphism.";
RL J. Clin. Invest. 88:1005-1013(1991).
CC -!- FUNCTION: COATS THE EPITHELIA OF THE INTESTINES, AIRWAYS, AND
CC OTHER MUCUS MEMBRANE-CONTAINING ORGANS. THOUGHT TO PROVIDE A
CC PROTECTIVE, LUBRICATING BARRIER AGAINST PARTICLES AND INFECTIOUS
CC AGENTS AT MUCOSAL SURFACES.
CC -!- SUBUNIT: MULTIMERIC.
CC -!- SUBCELLULAR LOCATION: SECRETED.
CC -!- TISSUE SPECIFICITY: COLON, SMALL INTESTINE, COLONIC TUMORS,
CC BRONCHUS, CERVIX AND GALL BLADDER.
CC -!- PTM: ALL CYSTEINE RESIDUES ARE INVOLVED IN INTRACHAIN OR
CC INTERCHAIN DISULFIDE BONDS (BY SIMILARITY).
CC -!- POLYMORPHISM: THE NUMBER OF REPEATS IS HIGHLY POLYMORPHIC AND
CC VARIES AMONG DIFFERENT ALLELES.
CC -!- SIMILARITY: THE N-TERMINAL DOMAIN SHOWS SOME SIMILARITY TO THAT
CC OF SILKWORM HEMOCYTIN.
CC -!- SIMILARITY: CONTAINS 2 VWFC DOMAINS.
CC -!- SIMILARITY: CONTAINS 1 C-TERMINAL CYSTINE KNOT-LIKE DOMAIN (CTCK).
CC ---

```
Qy    117 TSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKE----PASTTPKEPTPTTI 172
          ::|   |  ||     ||      | |    |    :: ::    ||  |    || |
Db    171 EASKKKQPVIKV----PPKQAVVKAGLASNNGKTADSSSSEDSDSPPAKKTAATKTPPT- 225

Qy    173 KSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTT-TKEPAPTT-----T 226
            | |  :|    |    :: :| :  :: |   |   :| |   :  | ||:     |
Db    226 --KPATAAKPQAKKTAGKKSSSREDSSDSSDEEQKTAKSKPKPDVYSAVPPPTSVSKKKT 283

Qy    227 KSAPTTPKEPAPT----------TPKKPAPTTPKEPAPTTP----KEPTPTTPKEPAPTT 272
          | | |  :|  :           | | |   | : | : |    |:   :|  |   ::
Db    284 LSQPGTKAKPESSDSSDSSDEEEQPAKKAKIVPAKAAASAPKPLAKKAETSTDSESDSSS 343

Qy    273 KEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTP----KEPAP----------------TT 311
          ::   :: |     |||| ||  |   | :||    |: ||                ||
Db    344 EDEKKSSVKLGVKAAPKK-APAAP--DAKSTPVAAAKKSAPAKKASSSSDSDSSSNEETT 400

Qy    312 TKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA- 370
          ||  : ||| : | | |   ||  |    | |   |  |  |  :| |: : ::    :
Db    401 TKPAAKTTPAKSAATPTSKTPTNGK---ATPTSKTPAKPGTPKTSTAKKDSSSSDSSDSS 457

Qy    371 ---PTTPKKPAPTTPKEPAPT-TPKEPAPTTTKKPAPTAPKEPA-PTTPKETA------- 418
             || |  | ||| : | | | | |   |  |  ||: | || | ||| :|
Db    458 SDEETTTKPAAKTTPAKSAATPTSKTP---TNSKATPTSKKTPAKPGTPKTSAAKKDSSS 514

Qy    419 -----PTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTP-KA 472
               :: :| ||   :: | ||| |||       | ||| :|        | ||| |
Db    515 SDSSDSSSDEKKTPA--KRAAKTTPAKPA-------AKTTPAKPA-------AKTTPAKP 558

Qy    473 AAPNTPKEPAPT----TPKEPAPTTPKEPAPTTPKETAPTTP-KGTAPTTLKEPAPTTPK 527
          || :|| :  ||    :  : : :: ::   :: |    ||| | |:        |
Db    559 AAKSTPGKQVPTKKESSSSDSSDSSSEDEKKSSAKPAVKTTPGKATS-------KPVVAS 611

Qy    528 KPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTT--PKEPAPTTPKGTAPT 585
          || | :      |  :|: ||    | | | | |   ||  |  ||:   ::   :
Db    612 KPVPAK------KASSSSDSDSSEEETTKTTKPLTKLSPAVKTLPPKKAESSSDSSSDSD 665

Qy    586 TLKEPAPTTPKKPAPKELAPTTTKGP--------TSTTSDKPAPTTPKETAPT--TPKEP 635
          : |:   | | ||  |   |  || |        : : ||  :    |   ||  :|
Db    666 SEKK---TKPAKPPAKSATPVNTKAPAQNKASKASCSDSDSSSEEEGKSKQPTGKSPAAK 722

Qy    636 APTTPKK-PAPTTPETPPPTTSEVSTPTTTKE-PTTIHKSPDESTPELSAEPTPKALENS 693
          |   ||| |     : |  ::|  |:    |: |     : |      :|:|||   :|
Db    723 ATAPPKKNPVAVNKDKPSSSSSSDSSGDDEKQKPKQAAAAKDVKQGAKAAKPTPKKAASS 782

Qy    694 PKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESK 753
            |    ::     :| : | ||  |:      |||:   || | :   :: :|   :
Db    783 SSE---DSSSDEDVSKAKKTNTAVSKSPV----TTPKAVPAAKKESSSESSDSEDEKQGG 835

Qy    754 ITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTK--KTITTTEIMNKPEETA-KPKD 810
            :||:: ::|                ||    :  :  :::|   |   |: | |
Db    836 KNTSTTKIANST-----------------PKAAAAECSEESSSSEDEGKANGTSGKRKR 877

Qy    811 RATNS---KATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTP-RKMTSTMPELNPT 866
          :| :   :| ||:          |    | | | |:| | :   || |::     |:||
Db    878 ESTGNAECEAVTPE--------NKKLKAKSPNTFPKVNKKELKNTPFRRVVEEDIEINP- 928

Qy    867 SRIAE 871
           |:|:
Db    929 -RMAD 932
```

Search completed: April 26, 2002, 16:18:10
Job time: 280 sec

```
Qy     767 QDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPT 826
           | |:   :   : ||    :: :   |:  :|||                  :|
Db    6877 VPEVPKKVE--EKRIILPK----EEEVLPVEVTEEPEE---------------EPISEE 6914

Qy     827 KAPKKPTSTKK-----PKTMPRVRK---PKT-TPTPRKMTSTMPELNPTSRIAEAMLQTT 877
           : |::| | ::     |  :| | |   |:  || |:|: :  |    : :| |  :
Db    6915 EIPEEPPSIEEVEEVAPPRVPEVIKKAVPEAPTPVPKKVEA--PPAKVSKKIPEEKVPVP 6972

Qy     878 TRPNQTPNSKLVEVN----------PKSEDAGGAEGET 905
           :  : | :|: ||           || |    |:| |
Db    6973 VQKKEAPPAKVPEVPKKVPEKKVLVPKKEAVPPAKGRT 7010


RESULT  14
T34513
hypothetical protein ZK783.1 - Caenorhabditis elegans
C;Species: Caenorhabditis elegans
C;Date: 29-Oct-1999 #sequence_revision 29-Oct-1999 #text_change 29-Oct-1999
C;Accession: T34513
R;Favello, A.; Vaudin, M.
submitted to the EMBL Data Library, August 1994
A;Description: The sequence of C. elegans cosmid ZK783.
A;Reference number: Z21536
A;Accession: T34513
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-3507 <FAV>
A;Cross-references: EMBL:U13646; PIDN:AAC24418.1; GSPDB:GN00021; CESP:ZK783.1
A;Experimental source: strain Bristol N2; clone ZK783
C;Genetics:
A;Gene: CESP:ZK783.1
A;Map position: 3
A;Introns: 14/1; 48/2; 84/1; 196/3; 303/1; 381/1; 586/1; 605/1; 1175/3; 1207/1; 1409/2;
 3504/1


  Query Match             12.6%;  Score 629;  DB 2;  Length 3507;
  Best Local Similarity   24.1%;  Pred. No. 1.7e-19;
  Matches  252;  Conservative  158;  Mismatches  416;  Indels  218;  Gaps   45;

Qy      10 KKKPTPKPPVVDEAGSGLDN-GDFKVTT----PDTSTTQHNKVSTSPKITTA-------K 57
           ||:|  :   :|| |   | | | ||       || |      :  | :||:       |
Db    1942 KKQPNREKIEIDEENSSSSNSGQEKPTTKGIVSSTSATSSESTTAEPHVTTSISSTTSTK 2001

Qy      58 PINPRPSLPPN---SDTSKETSLTVNKETTV-ETKETTTTNKQTSTDGKEKTTSAKETQS 113
            :    | | |   |  | | | : :| ||  ||  ::| :: :|::    :: |  |:
Db    2002 DMTSSKS-PENVTMSSESPEVSTSSSKSTTASETTVSSTPSESSSSEAPLTSSPATTTEV 2060

Qy     114 IEKTSAKDLAP----TSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTP 169
           | ::| |   |    :|::  | : |:   |:    ::|  |: || :   || |:
Db    2061 ITESSVKSTTPKEESSSEITVKLSSKSPEVTESSVKSSPSTPS-TTSQSVTSTVPETSKS 2119

Qy     170 TTIKS-APTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKS 228
           | : | || |   |    | |: | |   | :|| :  |||   :  : | :     |
Db    2120 TVLSSEAPVTSTSPTEVHT-SSETKPSLSASSTTGDTNSTTPSTSSLASVKSTSAPEGTS 2178

Qy     229 APTTPKEPAPTTPKKPAPTTPK-EPAPTTPKEPTPTTPKEPAPTTKEP-APTTPKEPAPT 286
           |   | : :  :|    |:|   :   ::  : : |:      :| || :  |      :
Db    2179 ASVAPVKLSSLSPDVSQPSTKTFDATESSTVQASETSSGTSVKSTSEPESHVTKLSITSS 2238

Qy     287 APKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTP--KEPAPTTTKSAPTTTKEPAPTTTK 344
           |   | |  :  || |:     :|::|: |||  :   |  : |   || ||   ::
Db    2239 NPSSSVPVTSPKSTPTVPE-----STEQPTSTTPSGQSLTPMNSNSEVLTTSEPHVLSSS 2293

Qy     345 SAP------TTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTT- 397
           :|      |||   | ::| |    |||  :  :     |:| : |   :|   : ||
Db    2294 LSPDVSQSSTTPNNLSESSTVE----TPKTSSEVSLNSEEPSTTEAPTTLSPDILSTTTN 2349

Qy     398 --------------------TKKPAPTAPK----------EPAPTTPKETAPTTPKKLTP 427
                              ::||  :||:          :|  | |:  : |  ||
Db    2350 NLSQSSTVSTEDRSEISSENSEKPT-SAPELVTSSVTHVASSSPDVPTES--SEPDDLTG 2406

Qy     428 TTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEP----------APTTPKAAAPN 476
           :  :
```

```
           :: | :   : ::   :||     |  |||| :|   |           : |||:::
Db    2407 SSTENIPEASSKQTISSTPTPDTTTASEEPTKSTSMSPDLSTTSNVLSESSTTPESSS-- 2464

Qy     477 TPKEPAPTT-----------PKEPAPT---------TPKEPAP-----TTPKETAPTTP 510
           | |  ::              | |  |          |   |:|     ||  ||:
Db    2465 --KSPVSSSTEGISVVTSTEFSKVPESTISSVLEEDLTKTTPSPILEETTTASETSEPLT 2522

Qy     511 KGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPA---------PT-------- 553
           : :   :::    ||  :  |||   |||   :|  | :||          ||
Db    2523 EDSLTVSVRIHELTTSSENVPKESESTTTSSESSKPSQEPAGILTSTVVVPTSSVSLITA 2582

Qy     554 ---------TP--KGTAP--TTPK-------EPAPTTPKEPAPTTPKGTAPTTLKEPAPT 593
                   ||  :|  |  |:||         |:  |  ||: :| : |  ||:    ||
Db    2583 SEIEAITSNTPFKQGRTPITTSPKSLVKSTTSPSTVTSSEPSESTKRTTVSTTVSTTTPT 2642

Qy     594 TPKKPAPKELAPTTTKGPT-STTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETPP 652
           :   :        || |||    ||||| :|: | |   : |: |         |
Db    2643 EETTTSESLILTAAPSKPTESTTESSEAPTTPAKTSETKPSNVSSTSRKSTENVETSTSQ 2702

Qy     653 PTTSEVSTPTTTKEPTTIHKSPDESTP--ELSAEPTPKALE--NSPKEPGVPTTKTPAAT 708
           : | || ::|         |: : |   :|:| :   ||  :|   |    | : |:
Db    2703 SGSLESSTMSSTS------SEPETNAPAVTVSSEASSTTLEENSSTSSP----TSSEASV 2752

Qy     709 K-----PEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTT---- 759
           |     ||  |:     : |     |  | : :  :| :| : : :| :|  :||
Db    2753 KLSSLFPESITSEAVTVSSR----APAEITMSSESHREISTVSSEPSEPEIPLSTTVSPN 2808

Qy     760 ----------QVTSTTTQDTTP-FKITT-----LKTTTLAPKVTTTKKTIITTTEIMNKP 802
                    : |: |  :||  :: |     |  :   |     :: || : :   |
Db    2809 VVTASSIPSEEPILSSVTSSSTPRVRLITGTPDDLIVSVTVPSHGNRRQNITASSV---P 2865

Qy     803 EETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPK---TTPTPRKMTST 859
           :  |    : |  |||:|   |    || :|   :  | :: |    |||
Db    2866 SNSTSPIILPSES-LTTPQPPPTTTTTAKPATTSGKRGPPSIQPPAEMFTTPAP------ 2918

Qy     860 MPELN-----PTSRIAEAMLQTTT 878
           | |       |::  | :  |||
Db    2919 PPPSNGGYGEETNQEEEQVTSTTT 2942


RESULT  15
I51618
nucleolar phosphoprotein - African clawed frog
C;Species: Xenopus laevis (African clawed frog)
C;Date: 13-Sep-1996 #sequence_revision 13-Sep-1996 #text_change 24-Sep-1999
C;Accession: I51618; S57757
R;Cairns, C.; McStay, B.
J. Cell Sci. 108, 3339-3347, 1995
A;Title: Identification and cDNA cloning of a Xenopus nucleolar phosphoprotein, xNopp
A;Reference number: I51618; MUID:96019267
A;Accession: I51618
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: mRNA
A;Residues: 1-990 <CAI>
A;Cross-references: EMBL:X88927; NID:g895920; PIDN:CAA61368.1; PID:g895921
C;Genetics:
A;Gene: xNopp180
C;Superfamily: nucleolus-cytoplasm shuttle phosphoprotein
C;Keywords: phosphoprotein


  Query Match             12.5%;  Score 626.5;  DB 2;  Length 990;
  Best Local Similarity   26.2%;  Pred. No. 7e-20;
  Matches  253;  Conservative  132;  Mismatches  393;  Indels  187;  Gaps   45;

Qy       3 DNKKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPR 62
           | || :      ||     |: | |:   :   |      |       |    | : |:
Db      59 DAKKRKPPANGLPKKKSAKESSSSEDSSSEEDEPPAKKRAQ----PAGGKKPVVKAVQPK 114

Qy      63 PSLPPNSDTSKETSLTVNKETTVE--TKETTTTNK----QTSTDGKEKTTSAKETQSIEK 116
             :  : |:| |:    ::| | :   |    | |    :| |  | |::|: |: | |
Db     115 KAKSSSEDSSDESD---SEEETKKPPAKRPAQTPKVAAVKTPTQKKAKSSSS-ESSSSED 170
```

```
Qy    132 PTPKAETTTKGPALTTP---KEPTPT-TPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTT 187
Db     69 PPPIQKPPTYSPPIYPPPIQKPPTPTYSPPIYPPPIQKPPTPTYSPPIYPPPIQKPPTPT 128

Qy    188 KSAPTTPKEPAPTTTKEPAP--TTPKEPAPTTTKEPAPTTTKSAPTTP----KEPAPT-- 239
Db    129 YSPPIYP----PPIQKPPTPSYSPPVKPPPV---QMPPTPTYSPPIKPPPVHKPPTPTYS 181

Qy    240 TPKKP---APTTPKEPAPTTP----KEPTP--TTPKEPAPTTKEPAPT-TPKEPAPTAPK 289
Db    182 PPIKPPVHKPPTPIYSPPIKPPPVHKPPTPIYSPPIKPPPVHKPPTPTYSPPVKPPPVHK 241

Qy    290 KPAPT-TPKEPAPTTPKEPAP---TTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKS 345
Db    242 PPTPIYSPPIKPPPVHKPPTPIYSPPVKPPPVQTPPTPIYSPPVKPPPVHKPPTPTYSPP 301

Qy    346 APTTPKEPSPTTTKEP---APTTPKEPAPT--TPKKPAPTTPKEPAPTTPKEPAPTTTKK 400
Db    302 VKSPPVQKPPTPTYSPPIKPPPVQKPPTPTYSPPIKPPPVKPPTPIYSPPVKP-PPVHKP 360

Qy    401 PAP--TAPKEPAPT----TPKETAPTTP----KKLTPTTPEKLAPTTPEKPAPTT---PE 447
Db    361 PTPIYSPPVKPPPVHKPPTPIYSPPVKPPPIQKPPTPTYSPPIKPPPLQKPPTPTYSPPI 420

Qy    448 ELAPTTPEEPTPTTPEEPAPT----TPKAAAP-NTPKEPAPTTPKEPAPTTPKEPAPTTP 502
Db    421 KLPPVKPPTPIYSPPVKPPPVHKPPTPIYSPPVKPPPVHKPPTPTYSPPIKPPPVKPPTP 480

Qy    503 KETAPTTPKGTAPTTLKEPAPT--TPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAP 560
Db    481 TYSPPVQP----PPVQKPPTPTYSPPVKPPPIQKPPT----PTYSPPIKPPPVKPPTPTY 532

Qy    561 TTPKEPAPTTPKEPAPT-TPKGTAPTTLKEPAPT--TPKKPAPKELAPTTTKGPTSTTSD 617
Db    533 SPPIKP-PPVHKPPTPTYSPPIKPPPIHKPPTPTYSPPIKPPPVHKPPTPTYSPPIKPPP 591

Qy    618 KPAPTTPKETAPTTPKEPAPTTPKKPAPT-TPETPPPTTSEVSTPTTTK--EPTTIHKSP 674
Db    592 VHKPPTPTYSPPIKP----PPVHKPPTPTYSPPIKPPPVHKPPTPTYSPPIKPPPVHKPP 647

Qy    675 DESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTA 734
Db    648 ---TPTYSPPIKPPPVQKPPTPTYSPPVKPPPVQLPP-TPTYSPPVKPPPVQVPPTPTYS 703

Qy    735 APKMTKETATTTEKTTESKITATTTQVTSTTTQDTTP 771
Db    704 PPVKPPPVQVPPTPTYSPPIKPPPVQVPPTPTTPSPP 740
```

RESULT 13
I38346
elastic titin - human (fragment)
C;Species: Homo sapiens (man)
C;Date: 29-May-1998 #sequence_revision 29-May-1998 #text_change 21-Jul-2000
C;Accession: I38346
R;Labeit, S.; Kolmerer, B.
Science 270, 293-296, 1995
A;Title: Titins: giant proteins in charge of muscle ultrastructure and elasticity.
A;Reference number: A57430; MUID:96026330
A;Accession: I38346
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: mRNA
A;Residues: 1-7962 <RES>
A;Cross-references: EMBL:X90569; NID:g1017426; PIDN:CAA62189.1; PID:g1017427
C;Genetics:
A;Gene: GDB:TTN
A;Cross-references: GDB:127867; OMIM:188840
A;Map position: 2q31-2q31

Query Match 12.6%; Score 632; DB 2; Length 7962;
Best Local Similarity 23.3%; Pred. No. 2.7e-19;
Matches 288; Conservative 102; Mismatches 442; Indels 406; Gaps 58;

```
Qy     10 KKKPTP----KPPV---VDEAGSGLDNGDFKVT-----------TPDTSTTQHNKVSTS 50
Db   5837 KKAPPPPKVIKKPVIEKIEKTSRRMEEEKVQVTKVPEVSKKIVPQKPSRTPVQEEVIEVK 5896

Qy     51 -PKITTAKPINPRPSLPPNSDTSKETSLT---VNKETTVETKETTTTNKQTSTDGKEKTT 106
Db   5897 VPAVHTKKMVISEEKMFFASHTEEEVSVTVPEVQKEIVTEEKIHVAVSKRVE-------P 5949

Qy    107 SAKETQSIEKTSAKDLAPT--SKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPAS--- 161
Db   5950 PPKVPELPEKPAPEEVAPVPIPKKVEPPAPKVPEVPKKP-VPEEKKPVPVPKKEPAAPPK 6008

Qy    162 --TTPKEPTPTTIKSAPTTPKEPAP-----------TTTKSAPTTPKEPAPTTTKEPA- 206
Db   6009 VPEVPKKPVPEEKIPVPVAKKKEAPPAKVPEVQKGVVTEEKITIVTQREESP----PPAV 6064

Qy    207 PTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTP-KEPAPTTPKEPTPTT- 264
Db   6065 PEIPKKKVPEERK-PVPRKEEEVPPPPKVPA--LPKKPVPEEKVAVPVPVAKKAPPPRAE 6121

Qy    265 ----------------------------------------------------------P 265
Db   6122 VSKKTVVEEKRFVAEEKLSFAVPQRVEVTRHEVSAEEEWSYSEEEEGVSISVYREEEREE 6181

Qy    266 KEPAPTTK-----------------------EPAPTT----------PKEPA------ 284
Db   6182 EEEAEVTEYEVMEEPEEYVVEEKLHIISKRVEAEPAEVTERQEKKIVLKPKIPAKIEEPP 6241

Qy    285 ----PTAPKK-------PAPTTPKE--PAPTTPKEP-APTTTKEPSPTTPKE----PAPT 326
Db   6242 PAKVPEAPKKIVPEKKVPAPVPKKEKVPPPKVPEEPKKPVPEKKVPPKVIKMEEPLPAKV 6301

Qy    327 TTKSAPTTTKEP--APTTTKSAP---TTPKEPSPTTTKE------------PAPTT--- 365
Db   6302 TEKHMQITQEEKVLVAVTKKEAPPKARVPEEPKRAVPEEKVLKLKPKREEEPPAKVTEFR 6361

Qy    366 -------------PK-EPAP---------------------------------------- 371
Db   6362 KRVVKEEKVSIEAPKREPQPIKEVTIMEEKERAYTLEEEAVSVQREEEYEEYEEYDYKEF 6421

Qy    372 ------------------------------TTPKKPAPT--TPKEPAPTTPKEPAPTTTK 399
Db   6422 EEYEPTEEYDQYEEYEEREYERYEEHEEYITEPEKPIPVKPVPEEPVPTKPKAPPAKVLK 6481

Qy    400 KPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEK-------LAPTTPEKPAPTTPEELAPT 452
Db   6482 KAVP-EEKVPVP-IPKKLKPPPPK--VPEEPKKVFEEKIHISITKREKEQVTEP---AAK 6534

Qy    453 TPEEPTPTTPEEPAPTTPKAAAP--NTPKEPAPTTPKEPA------------------- 490
Db   6535 VPMKPKRVVAEEKVPVPRKEVAPPVRVPEVPKELEPEEVAFEEEVVTHVEEYLVEEEEEY 6594

Qy    491 ------PTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTS 544
Db   6595 IHEEEEFITEEEVVPVIPVK-VPEVPRKPVPEE-KKPVPVPKKKEAPPAKVPEVPKKPEE 6652

Qy    545 TT------SDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPA--PTTPK 596
Db   6653 KVPVLIPKKEKP----PPAKVPEVPKKPVP-EEKVPVPVPKKVEAP-----PAKVPEVPK 6702

Qy    597 KPAPKEL----APTTTKGPTSTTSDKPAPTTPKETAPT-TPKE---PAPTTPKKPAPTTP 648
Db   6703 KPVPEKKVPVPAPKKVEAPPAKVPEVPKKLIPEEKKPTPVPKKVEAPPPKVPKKREPVPV 6762

Qy    649 ETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAAT 708
Db   6763 PVALPQEEEVLFEEEIVPEEEVLPEEEEVLPE-EEEVLPEEEEVLPEEEEIPPEEEEVPP 6821

Qy    709 KPEMTTTAKDKTTERDL--RTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTT 766
Db   6822 EEEYVPEEEEEFVPEEEVLPEVKPKVPVPAP-----VPEIKKKVTEKKVVIPKKEEAPPAK 6876
```

```
Qy     360 E---PAPT-------TPK------EPAPTTPKKPAP-------------------TTPK 383
           |  ||       |||      ||   | | | |                       :  |
Db     442 SHVVPKPTKKGTVKVTPKLELSFDEPTEIT-KAPHPGKLLEKKTYHFVLSDNFARYSEAK 500

Qy     384 E---------------PAPTTPKEPAP----TTTKKPAPTA--PKEPAPTT---PKETAP 419
           |               ||| :| :     |||::|| |    |    |||    |  ||
Db     501 ENDDYNHLDYNHYREAKEPTTTEESSTTEEVTTTEEPANTGNPPTTENPTTTEQPTSTAE 560

Qy     420 TTPKKLTPTTPEKLA---PTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPN 476
           :|   |  || : :    ||| ||   |      ||| :|   |   |   || ||   :
Db     561 STTTALPFTTEQTVTTEEPTTAEKSTATQ----KPTTTQESVST---EKTSTTKKA---S 610

Qy     477 TPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAP 536
           | :|  |||  ||  ||    : || | | |     :  ||  |   ||
Db     611 TTEE--PTTTDEPTTTT---ESSTTGKATTPELSTTSEETTTTELKITTE--------GS 657

Qy     537 TTTKEPTSTTSDKPAP----TTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAP 592
           |||:|||:|     |     || : |  ||   |  |: ||    |      |::
Db     658 TTTEEPTTTAIFAEASTGIITTDEETTSTTSTTPEITSTKE--IVTESAITQTSVSVVES 715

Qy     593 TTP----------------------KKPAPKELAPTTTKGP----------------TS 613
           :||                     ||   ||   |:| |                  |:
Db     716 STPRQLPERWKAIVNKFKHNLEVLKEKKRLLKEKESTSTTGSDSSETTTVVAENIDEVTT 775

Qy     614 TTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPE--TPPPTTSE-VSTPTTTKEPTTI 670
           |  :|   |||  |  :| :|   ||      || :  |  ||||  :| ||| ||:|
Db     776 TEKEKVVQTTPITTEKSTTQEETTTTTTTTEKTTSKTTTEKPTTSESATTETTTSEPST- 834

Qy     671 HKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPE 730
               |||                    |  | :||  | :| |:  ||  :   | |
Db     835 ----TEST---------------------TVDTSSATTEESSTAAETTTTSAE---TSE 865

Qy     731 TTTA-----------------------------------APKMTKETATTTEKTTE 751
           |||:                                  ||  | |    | |:
Db     866 TTTSESAAFITGESPENTALQSSSQKSEENESSAEKPGARRDFVPKKHKTTVKPAETTSA 925

Qy     752 SKITATTTQ-VTST---TTQDTTPFKITTLKTTTLAPKVT---TTKKTITTTEIMNKPEE 804
           : |||: :|:|   || :||| : |||    |    ||     : || | |:::|
Db     926 VAASTTTTEPITTTEKSTTLETTPIEATTLNEVTGPAFVTGAPVDETTINTLELLSK--- 982

Qy     805 TAKPKDRATNSKATTPKPQ----------------KPTKAPKKPTSTKKPKTMPRVRKP 847
                 |:: : |||                   ||||  ||
Db     983 -------INNTQISQPKPTDISKTDALSSLISGLIGSFTKAPMAPTI------------- 1022

Qy     848 KTTPTPRKMTSTMPELN-----------PTSRIAEA 872
           ||     :|:|   ||            ||  |  |
Db    1023 HTTTDAAFVTATEASLNDGSDKKIIDEAQPTDEIRRA 1059
```

RESULT 11
T11622
extensin class 1 precursor - cowpea
C;Species: Vigna unguiculata (cowpea)
C;Date: 16-Jul-1999 #sequence_revision 16-Jul-1999 #text_change 17-Nov-2000
C;Accession: T11622; S54155
R;Arsenijevic-Maksimovic, I.; Broughton, W.J.; Krause, A.
Mol. Plant Microbe Interact. 10, 95-101, 1997
A;Title: Rhizobia modulate root-hair-specific expression of extensin genes.
A;Reference number: Z17301; MUID:97155574
A;Accession: T11622
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-489 <ARS>
A;Cross-references: EMBL:X91836; NID:g1015936; PIDN:CAA62943.1; PID:g1015937
A;Experimental source: sub_species Red caloona
R;Arsenijevic-Maksimovic, I.; Broughton, W.J.; Krause, A.
submitted to the EMBL Data Library, April 1995
A;Description: A class of root-hair specific extensins involved in rhizobium/legume inte
A;Reference number: S54155
A;Accession: S54155
A;Status: preliminary
A;Molecule type: mRNA
A;Residues: 326-489 <AR2>
A;Cross-references: EMBL:X86030; NID:g791149; PID:g791150
C;Genetics:
A;Gene: Ext26G
C;Superfamily: hydroxyproline-rich glycoprotein
C;Keywords: glycoprotein; hydroxyproline
F;1-23/Domain: signal sequence #status predicted <SIG>
F;24-489/Product: extensin class 1 #status predicted <MAT>

Query Match 12.6%; Score 633; DB 2; Length 489;
Best Local Similarity 31.2%; Pred. No. 2e-20;
Matches 149; Conservative 33; Mismatches 252; Indels 44; Gaps 7;

```
Qy     179 PKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAP 238
           ||:  |    : |   | | |     |:|: |          | |   |  |   | | |
Db      39 PKQTPPYYYNAPPYYYKSPPP-----PSPSPP----------PPPYVHKYPPYYYKSPPP 83

Qy     239 TTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTP---KEPAPTAPKKPAPTT 295
           :|  | |   | | | :|  | |   | | | :  |:|  |   | | | :| | |
Db      84 PSPSPPPPYVYKSPPPPSPSPPPPYYYKSPPPPS--PSPPPPYYYKSPPPPSPSPPPPYY 141

Qy     296 PKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSP 355
           | | | :|  | |   | | | :|  | |   || |              | :|  | |
Db     142 YKSPPPPSPSPPPPYVYKSPPPPSPSPPPPYYYKSPP-------------PPSPSPPPP 187

Qy     356 TTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPK 415
           | | | :|  | |   | | | :|  | |   | | | :   | |   | | | :|
Db     188 YYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYVYKSPPPPSPSPPPPYYYKSPPPPSPS 247

Qy     416 ETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAP 475
           |   |   | :|    |   : | | :|    |   : | | :|  | |   |: |
Db     248 PPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPP 307

Qy     476 NTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELA 535
           :|  | |   | | | :|  | |   |   | :|    |   | | | :|  | |
Db     308 PSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYVYKSPPPPSPSPPPPYYY- 366

Qy     536 PTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTP 595
           |  |   :   | | |   |  | :|  | |   | | | :|    |   | | | :|
Db     367 ----KSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSP 422

Qy     596 KKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETPPP 653
           |     |   | |   :   | |   |   | :|  | |   | | | :|  |||
Db     423 SPP-----PPYVYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPP 475
```

RESULT 12
C84672
hypothetical protein At2g27380 [imported] - Arabidopsis thaliana
C;Species: Arabidopsis thaliana (mouse-ear cress)
C;Date: 02-Feb-2001 #sequence_revision 02-Feb-2001 #text_change 02-Feb-2001
C;Accession: C84672
R;Lin, X.; Kaul, S.; Rounsley, S.D.; Shea, T.P.; Benito, M.I.; Town, C.D.; Fujii, C.Y
M.; Koo, H.; Moffat, K.S.; Cronin, L.A.; Shen, M.; VanAken, S.E.; Umayam, L.; Tallon,
euss, D.; Nierman, W.C.; White, O.; Eisen, J.A.; Salzberg, S.L.; Fraser, C.M.; Venter
Nature 402, 761-768, 1999
A;Title: Sequence and analysis of chromosome 2 of the plant Arabidopsis thaliana.
A;Reference number: A84420; MUID:20083487
A;Accession: C84672
A;Status: preliminary
A;Molecule type: DNA
A;Residues: 1-761 <STO>
A;Cross-references: GB:AE002093; NID:g5306260; PIDN:AAD41992.1; GSPDB:GN00139
C;Genetics:
A;Gene: At2g27380
A;Map position: 2

Query Match 12.6%; Score 632; DB 2; Length 761;
Best Local Similarity 31.1%; Pred. No. 3.2e-20;
Matches 217; Conservative 44; Mismatches 354; Indels 82; Gaps 33;

```
C;Genetics:
A;Gene: GDB:MUC1; PUM
A;Cross-references: GDB:120705; OMIM:158340
A;Map position: 1q21-1q23
A;Introns: 20/1; 62/3; 1165/3; 1184/2; 1230/1; 1270/3; 1320/3
C;Superfamily: polymorphic epithelial mucin
C;Keywords: alternative splicing; duplication; glycoprotein; phosphoprotein; polymorphis
F;1-1344/Product: mucin 1 precursor, splice form A #status predicted <PREA>
F;1-62/Region: mucin 1 amino-terminal non-repetitive
F;1-23/Domain: signal sequence #link PREA #status predicted <SIGA>
F;1-19,29-32/Domain: signal sequence #link PREB #status predicted <SIGB>
F;1-19,29-1344/Product: mucin 1 precursor, splice form B #status predicted <PREB>
F;1-19,29-212,1033-1344/Product: mucin 1 precursor, epithelial tumor antigen splice form
F;138-1017/Region: 20-residue repeats (GSTAPPAHGVTSAPDTRPAP)
F;1143-1344/Region: mucin 1 carboxyl-terminal non-repetitive
F;1245-1272/Domain: transmembrane #status predicted <TRM>
F;1046,1064,1118,1144,1222/Binding site: carbohydrate (Asn) (covalent) #status predicted
F;1213/Binding site: phosphate (Tyr) (covalent) #status predicted

  Query Match           13.3%;  Score 666.5;  DB 1;  Length 1344;
  Best Local Similarity  28.4%;  Pred. No. 1.9e-21;
  Matches  302;  Conservative   70;  Mismatches  432;  Indels  261;  Gaps   57;

Qy      77 LTVNKETTVETKETTTTNK--QTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 133
           |||   ||     | |     :|| | || ||| :  |:  :: |: :: || ||:  :
Db      15 LTVLTATTAPKPATVVTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 74

Qy     134 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTP----------KE 166
           |    :||:|    |   ||                | |     ||||          :
Db      75 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVTSAPDNK 134

Qy     167 PTPTT-------IKSAPTTPKEPAPTTT-------KSAPTTPKEPAPTTTKEP------A 206
           | | :         : ||| |   ||| :|        |||  |   ||| :|  |      |
Db     135 PAPGSTAPPAHGVTSAPDT--RPAPGSTAPPAHGVTSAPDT--RPAPGSTAPPAHGVTSA 190

Qy     207 PTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTP----------KKPAPTTPKEP---- 252
           | |   ||| :|  ||   | ||| |   |  | |           :||| :   |
Db     191 PDT--RPAPGSTAPPAHGVT-SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGV 247

Qy     253 --APTTPKEPTPTTP----------KEPAPTTKEP-------APTTPKEPAPTAPK---- 289
             || |   |  | |              ||| :  |       || |   |  |||
Db     248 TSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGV 307

Qy     290 KPAPTTPKEPAPTTPK----EPAPTTTKEPSPTTPKEPAPTT---TKSAPTTTKEPAPTT 342
           || |   |  | |        || |   |  | |      |:   |: || :|  ||
Db     308 TSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGV 367

Qy     343 TKSAPTTPKEPSPTTTKEP------APTTPKEPAPTTPK----KPAPTTPKEPAPTTPK- 391
           | ||| |   |:| :|  |      || |   |  | |       || |   |  | |
Db     368 T-SAPDT--RPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPA 424

Qy     392 ---EPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPE- 447
              || |   |  |||     |.:  :| |        |       || |   |  | | |
Db     425 HGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPA 484

Qy     448 ---ELAPTTPEEPTPTTP----------EEPAP--TTPKA----AAPNTPKEPAPTTP-- 486
              || |   |  | |              |||  | | |    :||:|   |  | |
Db     485 HGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPA 544

Qy     487 --------KEPAPTTPKEP------APTT---PKETAPTTPKGT-APTTLKEPAPTTP-- 526
                   |||  :   |      || |   |  |||     | || |   |  | |
Db     545 HGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPA 604

Qy     527 --------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKG----TAPTTPKEPAPTTP- 571
                  :|||   ||      ||    :|||  | |      :|| |   |  |  | | |
Db     605 HGVTSAPDTRPAPGSTAP-PAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPP 663

Qy     572 ---------KEPAP--TTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPT 606
                     |||  | |      :|| |   |  | |           :|||   ||
Db     664 AHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAP- 722

Qy     607 TTKGPTSTTSDKPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPET 650
                                                                  | ||
Db     723 PAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGS--T 780

Qy     651 PPPTTSEVSTPTTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTK 703
           ||     | | |   | :       :  :||       |  |    : ::|     | :
Db     781 APPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGST 840

Qy     704 TPAA--------TKPEMTTTAKDK---TTERDLRTTPETT-------TAAPKMTKETATT 745
           | |          |:|   :||       |:  | |  | :|       |:||    :|
Db     841 APPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAP----DTRPA 896

Qy     746 TEKTTESKITATTTQVTSTTTQDTTP--FKITTLKTTTLAPKVTTTKKTITTT--EIMNK 801
           |       |:   |       | |    :|:   |  ||  |       |:  :
Db     897 PGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPA 956

Qy     802 PEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMP 861
           |  || |    |::  | | |   : ||     |  | | |     | :|  |      :: |
Db     957 PGSTAPPAHGVTSAPDTRPAP--GSTAPPAHGVTSAPDTRP---APGSTAPPAHGVTSAP 1011

Qy     862 ELNPT-SRIAEAMLQTTTRPNQTP--NSKLVEVNPKSEDAGGAEG 903
           :  |      |      |: |: |   |    |:  :  :| | |
Db    1012 DTRPAPGSTAPPAHGVTSAPDNRPALGSTAPPVHNVTSASGSASG 1056


RESULT  10
T25697
hypothetical protein F16F9.2 - Caenorhabditis elegans
C;Species: Caenorhabditis elegans
C;Date: 15-Oct-1999 #sequence_revision 15-Oct-1999 #text_change 18-Feb-2000
C;Accession: T25697
R;Fulton, B.
submitted to the EMBL Data Library, August 1996
A;Description: The sequence of C. elegans cosmid F16F9.
A;Reference number: Z20071
A;Accession: T25697
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-1229 <FUL>
A;Cross-references: EMBL:U67956; PIDN:AAB07691.1; GSPDB:GN00028; CESP:F16F9.2
A;Experimental source: strain Bristol N2; clone F16F9
C;Genetics:
A;Gene: CESP:F16F9.2
A;Map position: X
A;Introns: 35/1; 361/1; 384/1; 482/2; 517/1; 971/1; 1021/2; 1179/3

  Query Match           13.2%;  Score 659.5;  DB 2;  Length 1229;
  Best Local Similarity  29.0%;  Pred. No. 3.5e-21;
  Matches  307;  Conservative   78;  Mismatches  339;  Indels  333;  Gaps   52;

Qy      33 KVTT-PDTSTTQ---HNKVSTSPKITTA-KPINPRPSLPPNSDT-------SKETSLTVN 80
           ||||  | |||       | ||:|:| |     ||  :    :  |       |  | |  :
Db     119 KVTTSTDASTTNAPTTGKDSTTPEIITGIVVINSKSESVTDMSTTRFSTTLSPTTELLTS 178

Qy      81 KETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTT 140
           || | |  :|:| :||| |    :  |  || :  : :    : |:   ||    : |||
Db     179 PETLVSTDSSTST-EQTSPDNTTEIASPMETNTTTEATTTSVEPSVSTLAS---EDETTV 234

Qy     141 KGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPKEPAPT 200
           | :|      ||     ||| :||| |            | :||| : |     ||
Db     235 TAIAEST------TTVIAEVSTTTEEPTTT-----------AESTTKKSTT----KAPA 272

Qy     201 TTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTP----KKPAPTTPKEPAPTT 256
           ||:|| ||| :|   |||:    ||| |  :| |   ||||     |  | | |  | |
Db     273 TTEEPTPTTTEE--VTTTEAETSTTTSSETSTEK---PTTPLIDNKIAGPATGK---PET 324

Qy     257 PKEP-TPTTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKE- 314
           | | |||      |  |        |   |  |  |  | ::    |   |  |||
Db     325 THFPVTGTTPN--FDTATETPFVAKSEDKMTLSKTAATETTQQTTEVT-DGPEKETTKNV 381

Qy     315 ----PSPTTPKEPAPTTTKSAP-------TTTKEPAPT----TTKSAPTTPKEPSPTTTK 359
               | | |     :|| :|        || |     |    :|:|| |       ||||
Db     382 SIEIPITTVPLVETTSTTSTASKESDGFHTTLKLKVTTADSDSTESATTVKPFNEETTTK 441
```

```
Db    338 KPAPPPPPQQLPKAAAAAAPTGTELKPATAPPHGSPRANSHTVTVTPPNVPRAAAATVP- 396

Qy    251 EPAPTTPKEPTPTT----PKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKE 306

Db    397 -TAGAVPKASTGTTPAAAPQQPVP---KAAPVTPPSPQQAVPRAATAAA----APVTPQQ 448

Qy    307 P---APTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAP---TTTKSAPTTPKEPSPTTTKE 360

Db    449 PVTKAATTTNATPPPQPIPKAATTTTATPVTPQQPIPKAGTDAAPPPAVPKAPSDGRAAT 508

Qy    361 P----APTTPKEPAPTTPKKPAPTTPKEPAP---------TTPKEPAPT-TTKKPAPTAP 406

Db    509 PGVPNAATDPQKPPPTPQSVPSAVTEPKPQPRAAPPPSNEATPAVPSPSPNLKSPLPTIP 568

Qy    407 K-------EPAPTTPK---ETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELA------ 450

Db    569 KPVPLMALTPQPVTAQMVTQLAATKPSPIVPKASPK-ALMTPPPPPPGLPRALAAAKLLG 627

Qy    451 -PTTP----------EEPTPTTP----EEPAPTTP---------KAAAPNTPKE------ 480

Db    628 LPSSPVASAMHAKVTPRPLPASPVPMAASPASLGPDAARVALATNAASPGAKPEAAGGNG 687

Qy    481 ----PAPTTPKEPAPTTPKEPAPTTPKETAPT------------TPKGTAPT------- 516

Db    688 TLMAPMGAANTQMAPIGAAGAAQTAPMGAAHTHVSPMGAGGATQMSPTGAANTHMSPIGA 747

Qy    517 ------------TLKEPAPTTPKKPAPKELAPTTTKEP--------TSTTSDKPAPTTP 555

Db    748 GGATQMSPMGAANTQMSPMGATTTQMSPMGAAATTQPSPMGAAATQVTATSAGNTMQVSP 807

Qy    556 KG--TAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEP--APTTPKKPAPKELAPTTTKGP 611

Db    808 MGAATPPQTPSVGAATTP-QPSPM----GAATTLMSPMGAATTPQ---PSPMGAVTTQPP 859

Qy    612 ----TSTTSDKP-APTTPKET---APTTPKEP---APTTPKKP-APTTPETPP------- 652

Db    860 PMAATNTTQPPPMAASTPQSTPMGAATTTQSPPMGATTTQSPPMGASTPQAPPTVAGSPT 919

Qy    653 PTTSEVSTPTTTKEPTTIHKSPDESTPEL--SAEPTPKALENSPKEPGVPTTKTPAATKP 710

Db    920 PPPPIPPSPTAQTSPQPMSKSPPPDPPKAPSAAAQTSPAAHVANASPGV-TAVSPA---P 975

Qy    711 EMTTTAKDKTTERDLRTTPETTTAAPKMT-KETATTTEKTTESKITATTTQVTSTTTQDT 769

Db    976 IGVTEASPSADGARLSPGPTAATDGPKASPAATADVTEAATD--VTAAATAVPA----EA 1029

Qy    770 TPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAP 829

Db   1030 AP---------TKAKRSSSSSSSSSSSSSSSSSSSSSSSSDSDSSSSSSESNPASPA--- 1077

Qy    830 KKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLV 889

Db   1078 -----------PAVGDGQQQMTPGAAQSVPP-------VTEAAVQ-------------- 1104

Qy    890 EVNPKSEDAGGAEGE 904

Db   1105 EAAAAAAAAAGAERE 1119
```

RESULT 9
A35175
mucin 1 precursor, repetitive splice form A [validated] - human
N;Alternate names: breast carcinoma-associated DF3 antigen; core protein KP39; episialin
ncreatic mucin; polymorphic epithelial mucin (PEM)
N;Contains: mucin 1 precursor, epithelial tumor antigen splice form; mucin 1 precursor,
C;Species: Homo sapiens (man)
C;Date: 20-Apr-2000 #sequence_revision 20-Apr-2000 #text_change 02-Jun-2000
C;Accession: A35175; B35175; A35886; A35887; S10572; S40293; A36735; PX0066; S10218; S51
R;Ligtenberg, M.J.L.; Vos, H.L.; Gennissen, A.M.C.; Hilkens, J.
J. Biol. Chem. 265, 5573-5578, 1990
A;Title: Episialin, a carcinoma-associated mucin, is generated by a polymorphic gene enc
A;Reference number: A35175; MUID:90202794
A;Accession: A35175
A;Molecule type: mRNA
A;Residues: 1-952,1033-1344 <LIG1>
A;Cross-references: GB:M32738; GB:J05288; NID:g182121; PIDN:AAA35804.1; PID:g182124;
A;Experimental source: splice form A
A;Note: GenBank entries HUMEPISIA1 and HUMEPISIA2 present only the amino-and carboxyl
A;Accession: B35175
A;Molecule type: mRNA
A;Residues: 1-19,29-952,1033-1344 <LIG2>
A;Cross-references: GB:M32739; GB:J05288; NID:g182126; PIDN:AAA35806.1; PID:g182129;
A;Experimental source: splice form B
A;Note: GenBank entries HUMEPISIB1 and HUMEPISIB2 present only the amino-and carboxyl
R;Gendler, S.J.; Lancaster, C.A.; Taylor-Papadimitriou, J.; Duhig, T.; Peat, N.; Burc
J. Biol. Chem. 265, 15286-15293, 1990
A;Title: Molecular cloning and expression of human tumor-associated polymorphic epith
A;Reference number: A35886; MUID:90368715
A;Accession: A35886
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 1-19,29-992,1033-1344 <GEN>
A;Cross-references: GB:J05581; NID:g188869; PIDN:AAA59876.1; PID:g188870
A;Note: GenBank entry HUMMUCAB includes one copy of the tandemly repeated sequence
R;Lan, M.S.; Batra, S.K.; Qi, W.N.; Metzgar, R.S.; Hollingsworth, M.A.
J. Biol. Chem. 265, 15294-15299, 1990
A;Title: Cloning and sequencing of a human pancreatic tumor mucin cDNA.
A;Reference number: A35887; MUID:90368716
A;Accession: A35887
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 1-19,29-1109,'S',1111-1339,'A',1341-1344 <LAN>
A;Cross-references: GB:J05582; NID:g189598; PIDN:AAA60019.1; PID:g189599
A;Note: GenBank entry HUMPANMU contains four fewer copies of the tandemly repeated se
R;Wreschner, D.H.; Hareuveni, M.; Tsarfaty, I.; Smorodinsky, N.; Horev, J.; Zaretsky,
Eur. J. Biochem. 189, 463-473, 1990
A;Title: Human epithelial tumor antigen cDNA sequences. Differential splicing may gen
A;Reference number: S10571; MUID:90276413
A;Accession: S10572
A;Molecule type: mRNA
A;Residues: 1-19,29-155,'P',157-175,'P',177-182,'A',184-212,1033-1344 <WRE>
A;Cross-references: EMBL:X52229; NID:g37053
R;Wreschner, D.H.
submitted to the EMBL Data Library, March 1990
A;Reference number: S40293
A;Accession: S40293
A;Molecule type: mRNA
A;Residues: 1-19,29-155,'P',157-175,'P',177-182,'A',184-212,1033-1037,'A',1039-1344 <
A;Cross-references: EMBL:X52229; NID:g37053; PIDN:CAA36478.1; PID:g37054
R;Abe, M.; Siddiqui, J.; Kufe, D.
Biochem. Biophys. Res. Commun. 165, 644-649, 1989
A;Title: Sequence analysis of the 5' region of the human DF3 breast carcinoma-associa
A;Reference number: A36735; MUID:90088473
A;Accession: A36735
A;Molecule type: mRNA
A;Residues: 1-142,'Q',144-162,'Q',164-168 <ABE>
A;Cross-references: EMBL:M31823; NID:g181542; PIDN:AAA35757.1; PID:g181543
R;Masuzawa, Y.; Miyauchi, T.; Hamanoue, M.; Ando, S.; Yoshida, J.; Takao, S.; Shimazu
J. Biochem. 112, 609-615, 1992
A;Title: A novel core protein as well as polymorphic epithelial mucin carry peanut ag
A;Reference number: JX0235; MUID:93123189
A;Accession: PX0066
A;Molecule type: mRNA
A;Residues: 998-1011,'ES',1014-1017;1018-1032,'T',1034-1037;1038-1057 <MAS>
A;Experimental source: gastric carcinoma cell
R;Zrihan-Licht, S.; Baruch, A.; Elroy-Stein, O.; Keydar, I.; Wreschner, D.H.
FEBS Lett. 356, 130-136, 1994
A;Title: Tyrosine phosphorylation of the MUC1 breast cancer membrane proteins cytokin
A;Reference number: S51026; MUID:95080414
A;Contents: annotation
A;Note: undetermined tyrosine residues in the carboxyl-terminal non-repetitive region
C;Comment: This protein is length polymorphic. Individuals may have between 21 and 12
 partial repeats. The repeat shown is defined by SmaI nuclease sites.
C;Comment: Serine and threonine residues in the tandem repeat domain are extensively
C;Comment: For an alternative splice form without a tandem repeat domain, see PIR:S48
```
```

```
A;Accession: A26877
A;Molecule type: DNA
A;Residues: 1-242 <YAM>
A;Cross-references: EMBL:M16164; NID:g172522; PIDN:AAA35014.1; PID:g172525
A;Accession: B26877
A;Molecule type: DNA
A;Residues: 762-1331 <YA2>
A;Cross-references: EMBL:M16165; NID:g172523; PIDN:AAA35015.1; PID:g172526
R;Pardo, J.M.; Ianez, E.; Zalacain, M.; Claros, M.G.; Jimenez, A.
FEBS Lett. 239, 179-184, 1988
A;Title: Similar short elements in the 5' regions of the STA2 and SGA genes from Sacchar
A;Reference number: S27281; MUID:89031230
A;Accession: S27281
A;Molecule type: DNA
A;Residues: 1-31 <PAR>
A;Cross-references: EMBL:X13857; NID:g4551; PIDN:CAA32069.1; PID:g4552
R;Lambrechts, M.G.; Bauer, F.F.; Marmur, J.; Pretorius, I.S.
Proc. Natl. Acad. Sci. U.S.A. 93, 8419-8424, 1996
A;Title: Mucl, a mucin-like protein that is regulated by Mss10, is critical for pseudohy
A;Reference number: JC6123; MUID:96323237
A;Accession: JC6123
A;Status: nucleic acid sequence not shown
A;Molecule type: DNA
A;Residues: 1-1367 <LAM>
A;Cross-references: GB:U30626; NID:g1304386; PIDN:AAC49609.1; PID:g1304387
C;Genetics:
A;Gene: SGD:MUC1; STA2; MAL5; DEX2; SGD:S0001458
A;Cross-references: MIPS:YIR019c; SGD:S0001458
A;Map position: 9R
C;Superfamily: yeast glucan 1,4-alpha-glucosidase homolog; glucan 1,4-alpha-glucosidase
C;Keywords: glycosidase; hydrolase; polysaccharide degradation; transmembrane protein
F;5-21/Domain: transmembrane #status predicted <TM1>
F;1350-1366/Domain: transmembrane #status predicted <TM2>

  Query Match             15.2%;  Score 762.5;  DB 1;  Length 1367;
  Best Local Similarity   27.8%;  Pred. No. 1.7e-25;
  Matches  280;  Conservative  115;  Mismatches  441;  Indels  171;  Gaps   42;

Qy      35 TTPDTSTTQHNKVSTSPKITTAKPINP--------RPSLPPNSDTSKETSLTVNKETTVE 86
           :| ::|||  :   :|  :|  | |         :|: |  :  :||     : :||
Db     236 STSESSTTTSSTSESSTSSSTTAPATPTTTSCTKEKPTPPTTTSCTKEKPTPPHHDTTPC 295

Qy      87 TKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETT--TKGPA 144
           ||: |||:| | |   :|||:   | |   |:    ||       ||| : ||  :  |
Db     296 TKKKTTTSK-TCT---KKTTTPVPTPS-SSTTESSSAPV------PTPSSSTTESSSAPV 344

Qy     145 LTTPKE------PTPTTPKEPASTTPKEPTPTTIKSAP----TTPKEPAPTTTKSAPTTP 194
           ::  |      |||::    :|: |   : |  |||    ||    ||  | |: ||
Db     345 TSSTTESSSAPVPTPSSSTTESSSAPVTSSTTESSSAPVTSSTTESSSAPVPTPSSSTTE 404

Qy     195 KEPAP----TTTKEPAP----TTPKEPAP----TTTKEPAP----TTTKSAPTTPKEPAP 238
           ||    ||   ||    ||    ||    ||    ||  |:   |   :
Db     405 SSSAPVTSSTTESSSAPVTSSTTESSSAPVTSSTTESSSAPVTSSTTESSSAPVPTPSSS 464

Qy     239 TTPKKPAPTTPKEPAPTTPKEPTP---TTPKEPAPTTKE-------PAPT----TPKEPA 284
           ||    || |      ::   |||   ||    || |         | ||    | :  :
Db     465 TTESSSAPVTSSTTESSSAPVPTPSSSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSS 524

Qy     285 PTAPKKPAPTTPKEPAPTTPKE--------PAPTTTKEPSPTTPKEPAPTTTKSAPTTTK 336
           ||   : ||    || |           | |:::   | :||   : | : |||  |
Db     525 APAPTPSSSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSSTPVTSSTTESSSAPVPT- 583

Qy     337 EPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTT------PKEPAPTTP 390
           |: :||:|:      || :||:  :   |   : ||    || |           | ||
Db     584 -PSSSTTESSSAPVPTPSSSTTESSSAPAPTPSSSTTESSSAPVTSSTTESSSAPVPT-- 640

Qy     391 KEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAP----TTPEKPAPTTP 446
           |: :||:  :   |   : ||   :|| |   : ||    ||    || | |
Db     641 --PSSSTTESSSAPVPTPSSSTTESSSAP-VPTPSSSTTESSSAPVTSSTTESSSAPVTS 697

Qy     447 EELAPTTPEEPTP---TTPEEPAPT-TPKAAAPNTPKEPAP-----TTPKEPAPTT---- 493
             ::   |||   ||   ||  || ::   :   | |     ||    || |
```

```
Db     698 STTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESSSAPVTSSTT 757

Qy     494 -------PKEPAPTTPKETAPT-TPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTST 545
                 |   : ||   :||  ||  :   :  | ||         :|| |   :|
Db     758 ESSSAPVPTPSSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESSVAPVPTPSSSSN 817

Qy     546 TSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAP 605
           :     :||  ::  :   | ||     : ||   :|| :      :    | |   :
Db     818 ITSSAPSSTPFSSSTESSSVPVPT---PSSSTTESSSAPVSSSTTESSVAPVPTPSSSSN 874

Qy     606 TTTKGP-----TSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPET--PPPTTSEV 658
             |:  |     :|||      ||     |:: | |   |    : ||  |  |  ||: |
Db     875 ITSSAPSSIPFSSTTESFSTGTT---VTPSSSKYPGSQTETSVSSTTETTIVPTKTTTSV 931

Qy     659 STPTTTKEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTT---- 714
           :||:||   ||:  :   |  | ::  :|| :  :     ||| | : |    ||
Db     932 TTPSTTTITTTVCSTGTNSAGETTSGCSPKTVTTT-----VPTTTTTSVTTSSTTTITTT 986

Qy     715 ---TAKDKTTERDLRTTPET-TTAAPKMTKETATTTEKTTESKITATTTQVTST--TTQD 768
             |  :   |     :|:| ||  |  | : | :| || |  |  || |::|  ||:
Db     987 VCSTGTNSAGETTSGCSPKTITTTVPCSTSPSETASESTTTSPTTPVTTVVSTTVVTTEY 1046

Qy     769 TTPFK------------ITTLKTTTLA--PKVTT----TKKTITTTEI---MNKPEET- 805
           :|  |              | |   ||:|  | |||     |  |||||       |   ||
Db    1047 STSTKPGGEITTTFVTKNIPTTYLTTIAPTPSVTTVTNFTPTTITTTVCSTGTNSAGETT 1106

Qy     806 --AKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPEL 863
             ||   |    :|   :  |:|    |:      :       |    :  |    :
Db    1107 SGCSPKTVTTTVPCSTGTGEYTTEATTLVTTAVTTTVVTTESSTGTNSAGKTTTGYTTKS 1166

Qy     864 NPTSRIAEAMLQTTTRPNQ--TPNSKLVEVNPKSED---AGGAEGET 905
             ||: :      ||  |:   || :  |      : :   |  | |||
Db    1167 VPTTYV------TTLAPSAPVTPATNAVPTTITTTECSAATNAAGET 1207


RESULT   8
T18535
high molecular mass nuclear antigen - chicken (fragment)
C;Species: Gallus gallus (chicken)
C;Date: 15-Oct-1999 #sequence_revision 15-Oct-1999 #text_change 15-Oct-1999
C;Accession: T18535
R;Shimada, K.; Harata, M.; Mizuno, S.
J. Cell Sci. 110, 3031-3041, 1997
A;Title: A nuclear matrix-associated high molecular mass nuclear antigen, HMNA, of ch
A;Reference number: Z18955; MUID:9803440
A;Accession: T18535
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: mRNA
A;Residues: 1-1151 <SHI>
A;Cross-references: EMBL:D88440; NID:d1177138; PID:d1025045; PIDN:BAA24137.1

  Query Match             13.4%;  Score 671;  DB 2;  Length 1151;
  Best Local Similarity   25.9%;  Pred. No. 1.1e-21;
  Matches  284;  Conservative  100;  Mismatches  419;  Indels  292;  Gaps   55;

Qy      14 TPKP-------PVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTS--PKITTAKPINPRPS 64
           || |       |   :  :|   ||       | :   | :    || |||     ||
Db     113 TPPPSQGPAGTPPPSQGAAGAPKGDGTAQPSGTKSGADGKPAAQDVPKATTA-ATEARP- 170

Qy      65 LPPNSDTSKETSLTVNKETTVETKETTTTNK--QTSTDGKEKTTSAKETQSIEKTSAKDL 122
             :   | || | |   | |  :    : :||    | :   :||  | :  ::
Db     171 -------ASAASPTVPKATAEATAVTAASQSAPKAATDAAAVTAA---SQSAPKATV-EV 219

Qy     123 APTSKVLAKPTPKAETTTKGPALTTPKEPTP-TTPKEPASTTPKEP----TPTTIKSAPT 177
           | :  :||          |  |   :| | |:|  | |:   :|    :||  |: |
Db     220 KPAAAAVAKEAKAVTAAAAAPKATAEAKPAPVTSPTIPCSSAEAKPLTAASPTASKA--T 277

Qy     178 TPKEPAPTT-----------TKSAPTTPKEPAPTTTK---------------------- 203
           :| | |               | ||:     | | ||
Db     278 AEAKPVPATASLMATKVTAEAKPAPSPSVPKATTDTKAVTATAPKAGPDVKPAVAVCAEA 337

Qy     204 EPAPTTPKEPAPTTTKEPAPTTTKSAPTT------PKEPA-------PTTPKKPAPTTPK 250
```

```
Db    769 PPAPLSSPPPAPQVKSSPPPVQVSSP-PPAPKSSPPLAPVSSPPQVEKTSPPPAPLSSPP 827

Qy    412 TTPKETAP----TTP----KKLTPTTPEKLAPTTPEKPAP-----TTPEELAPTTPEEPT 458

Db    828 LAPKSSPPHVVVSSPPPVVKSSPPPAPVSSPPLTPKPASPPAHVSSPPEVVKPSTP--PA 885

Qy    459 PTTPEEPAPTTPKAAAPNTPKE-PAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTT 517

Db    886 PTTVISP-PSEPKSSPPPTPVSLPPPIVKSSPPPAMVSSP-PMTPKSSPPPVVVSSPPPT 943

Qy    518 LKE---PA-----PTTPKK---PAPKELAPTTTKE---PTSTTSDKPAPTTPKGTAPTTP 563

Db    944 VKSSPPPAPVSSPPATPKSSPPPAPVNLPPPEVKSSPPPTPVSSPPPA---PKSSPPPAP 1000

Qy    564 -KEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPT 622

Db   1001 MSSPPPPEVKSPPPPAPVSSPPPPVKSPPPPAPVSSPP---PPVKSPPPPAPVSSPPPPV 1057

Qy    623 -TPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPEL 681

Db   1058 KSPPPPAPISSPPPPVKSPPPPAPVS--SPPPPVKSPPPPAPVSSPPPPIKSPPPPAPVS 1115

Qy    682 SAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPET 731

Db   1116 SPPPAP------VKPPSLP-PPAPVSSPPPVVTPAPPKKEEQSLPPPAES 1158
```

RESULT 6
T30826
nascent polypeptide-associated complex alpha chain, muscle splice form gp220 - mouse
N;Alternate names: alpha-NAC protein
C;Species: Mus musculus (house mouse)
C;Date: 22-Oct-1999 #sequence_revision 22-Oct-1999 #text_change 24-Nov-1999
C;Accession: T30826
R;Yotov, W.V.; St-Arnaud, R.
Genes Dev. 10, 1763-1772, 1996
A;Title: Differential splicing-in of a proline-rich exon converts alphaNAC into a muscle
A;Reference number: Z20889; MUID:96312450
A;Accession: T30826
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-2187 <YOT>
A;Cross-references: EMBL:U48363; NID:g1666688; PID:g1666689; PIDN:AAB18732.1
C;Genetics:
A;Gene: Naca
A;Map position: 10
A;Introns: 24/1; 1996/1; 2024/3; 2050/3; 2099/3; 2142/3; 2183/3
A;Note: differential splicing converts alphaNAC into a tissue-specific DNA-binding activ
C;Keywords: alternative splicing; DNA binding; transcription factor

Query Match 15.5%; Score 776; DB 2; Length 2187;
Best Local Similarity 27.7%; Pred. No. 7.2e-26;
Matches 288; Conservative 122; Mismatches 389; Indels 240; Gaps 49;

```
Qy     34 VTTPDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTSKE---------TSLTVNKETT 84

Db    806 VQSPKVDPIMSDVTPTSPKKTSATAV-PK-------DTSATLSLKSVPAVTSLSPPKAPV 857

Qy     85 VETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPA 144

Db    858 APSNEATIVPTEIPTSLKNALAAATPKETLATSIPKVTSPS----PQKTPKSVSLKGAPA 913

Qy    145 LTT-----------------PKE-------PTPTTPKEPASTT---------PKEPTPTT 171

Db    914 MTSKKATEIAASKDVSPSQFPKEVPLLQHVPPTSPPKSPVSDTLSGALTSPPPKGP-PAT 972

Qy    172 IKSAPTTPKE-PAPTTTKSAPTTPK---------EPAPTTTKEPAPTTPKEPAPTTTKEP 221

Db    973 LAETPTYPKKSPKPAASKKTPATPSPEGVTAVPLEIPPCSKKAPKTAAPKESSATSSSKR 1032

Qy    222 APTT--TKSAPT--------------------TPKEPAPTTPKK-PAPTTPKEPAPTTP 257
```

```
Db   1033 APKTAVSKEIPSKGVTAVPLEISLPLKETSKSATPGEKSASSPKRSPKTAGPKE---TPP 1089

Qy    258 KEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTT--PKEPAP----TTPKEPAPTT 311

Db   1090 GGVTAVPPEISLPPKETPQNATPNESLAASSQKRSPKTSVPKETPPGGVTAMPLEIPSAP 1149

Qy    312 TKEPSPTTPK-----EPAPTTTKSAPTTTKEPAPTTT-KSAPTTPK--------EPSPTT 357

Db   1150 QKAPKTAVPKQIPTPEDAVTILAGSPLSPKKASKTAAPKEAPATPSVGVIAVSGEISPSP 1209

Qy    358 TKEPAPTTPKEPAPTTPKKPAPTT--PKE-----------PAPTTPKEPAP-------T 396

Db   1210 KKTSKTAAPKENSATLPPKRSPKTAAPKETPATSSEGVTAVPSEISPSPPTPASKGVPVT 1269

Qy    397 TTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEK-PAPTTPEELAPTTPE 455

Db   1270 LTPKGAPNALAE-SPASPKKVPKTAAPEETSTTP------SPQKIPKVAGPKEASATPPS 1322

Qy    456 EPTPTT--------PEEPAPTTPKAAAPNTPKEPAPTTPKE-PAPTTPKEPAPTTPKETA 506

Db   1323 KKTPKTAVPKETSAPSEGVTAVPLEIPPSPRKAPKTAAPKETPAPS--PEGATTAPVQIP 1380

Qy    507 PTTPKGTAPTTLKEPAPTTP-------------------KKPAPKELAPTTTKEPTS-- 544

Db   1381 PSPRKGSKKAGSKE-TPTTPSPEGVTAAPLEIPISSKKTSKMASPKETLVTPSSKKLSQT 1439

Qy    545 ----TTSDKPAPTTPKGTAPTTPKEPAPTTPKE-PAPTTPKGTAPTTLKEPAPTTPKKPA 599

Db   1440 VGPKETSLEGATAVPLEIPPSHKKAPKTVDPKQVPLTPSPK-DAPTTLAE-SPSSPKK-A 1496

Qy    600 PKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETP-----PPT 654

Db   1497 PKTAAPPSER-VTTVPPEKPA-TPQKASGTTASKVPVPAETQEVAVSSRETPVTPAVPPV 1554

Qy    655 TSEVSTPTTTK------EPTTIHKSPDESTPELSAEPTPKALENSPKE-PGVPTTKTPAA 707

Db   1555 KNPSSHKKTSKTIELKEAPATLPPSPTKSPKIPSSKKAPRT--SAPKEFPASPSIK-PVT 1611

Qy    708 TKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQ 767

Db   1612 T--SLAQTAPPSLQKAPSTTIPKENLAAPAV----LPVSSKSPAAPARASASLSPATAAP 1665

Qy    768 DTTPPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTK 827

Db   1666 QTAPKEATTIPSCKKAAATETPIETSTAPSLEGAPKETSE----TSVSKVLMSSP----- 1716

Qy    828 APKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSK 887

Db   1717 -PKKASSSKRASTLP-------------ATTLPSLKEASVLS---------PTATSSGK 1752

Qy    888 LVEVNPKSEDAGGAEGETP 906

Db   1753 DSHISPVS-DACSTGTTTP 1770
```

RESULT 7
S48478
glucan 1,4-alpha-glucosidase (EC 3.2.1.3) - yeast (Saccharomyces cerevisiae)
N;Alternate names: extracellular glucoamylase; mucin-like protein MUC1; protein YIR01
C;Species: Saccharomyces cerevisiae
C;Date: 10-Sep-1999 #sequence_revision 10-Sep-1999 #text_change 12-Nov-1999
C;Accession: S48478; A26877; B26877; S27281; JC6123
R;Rowley, K.
submitted to the EMBL Data Library, October 1994
A;Reference number: S48478
A;Accession: S48478
A;Molecule type: DNA
A;Residues: 1-1367 <ROW>
A;Cross-references: GB:Z47047; EMBL:Z38061; NID:g603997; PID:g763364; GSPDB:GN00009;
R;Yamashita, I.; Nakamura, M.; Fukui, S.
J. Bacteriol. 169, 2142-2149, 1987
A;Title: Gene fusion is a possible mechanism underlying the evolution of STA1.
A;Reference number: A91831; MUID:87194600

RESULT 4
T16251
hypothetical protein F35A5.1 - Caenorhabditis elegans
C;Species: Caenorhabditis elegans
C;Date: 20-Sep-1999 #sequence_revision 20-Sep-1999 #text_change 18-Feb-2000
C;Accession: T16251
R;Leimbach, D.
submitted to the EMBL Data Library, January 1996
A;Description: The sequence of C. elegans cosmid F35A5.
A;Reference number: Z18485
A;Accession: T16251
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-1274 <LEI>
A;Cross-references: EMBL:U46675; NID:g1166613; PID:g1166621; PIDN:AAB52641.1; GSPDB:GN0C
A;Experimental source: strain Bristol N2; clone F35A5
C;Genetics:
A;Gene: CESP:F35A5.1
A;Map position: X
A;Introns: 1272/2

Query Match 16.0%; Score 802; DB 2; Length 1274;
Best Local Similarity 28.0%; Pred. No. 3.5e-27;
Matches 290; Conservative 102; Mismatches 414; Indels 230; Gaps 56;

```
Qy      6 KNRTKK-KP-----TP-----KPPVVDEAGSGLDNGD----FKVTTPDTSTTQHNKVSTS 50
          || ||| ||      ||     :||| ::    |   |     |   | |    || |
Db    274 KNPTKKWKPPWEDETPVEEVKEPPVPEKKAPVLKKKDPAPAAKARDPSPSKAAPKKVEPS 333

Qy     51 PKITTAKPI-NPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAK 109
            :   |: ||     |  :   | :  | | :  | |      |:        ::
Db    334 SPVVPPTPVKNPVKKYKPPWEVDDEPAEEVKKPSAPEKKTPVLKRKEPEPSSTTPSSDPS 393

Qy    110 ETQSIEKTSAKDLAPTSKVLAKPTPKAE----TTTKGPA----------------LTTP 148
            ::      :| :|      :  |||:    |  | |                  :  |
Db    394 PKKAAPAVKPRDSSPKKATPLQADPKAQEVPPTPVKNPVKKYKPPWEVDDEDPVEEVKQP 453

Qy    149 KEPTPTTP----KEPASTTPKEP----TP-TTIKSAPTTPKEPAPTTTKSAPTTPKEPAP 199
          : |   ||    ||||:    :|    || |  |  |  |:: :|    : | : : ||
Db    454 EAPAKKTPVLKRKEPAAKDTAKPATSKTPETPEKKDPVKPRDSSPKKVAAKPDSAQAPA- 512

Qy    200 TTTKEPA------------------PTTPKEPAPTTTKEPAPT-------TTKSAPTTP 233
          | | |                    ||  |:      |:|||           ||  |
Db    513 TPVKNPVKKWRPPWEDDETPADDVSKPTDAKKTPSLAKKDPAPAKESLKPKADTKAPAKP 572

Qy    234 KEP-----APTTPKKPAPTTPKE----PAPTTPKEPTPTTPKEPAPTTKEPAPTTPK-EP 283
          ::|     ||| |:|  |   |:    || :  |||  : |::|:|    ||   || |
Db    573 RDPSPKKVAPTAPEKKTPVLAKKEPAGPADSKTKEPEKSKPRDPSPKKAVPAKPVPKTEV 632

Qy    284 APTAPKKPAP------TTPKEPAPTTPKEPAPTTTKEP----SPTTPKEPAPTTTKSAPT 333
          || | ||| |       | ||:  | :|  | ||  | |      |    : ||    | |
Db    633 APAAVKKPEPISKPKDTAPKKAEPNSPVVP-PTPVKNPVKKWKPPWEDDDAPAKPVSLPE 691

Qy    334 TTKEPAPTTTKSAPTTP-KEPSPTTTKEPAPTTPK--EPAPTTPKKPAP--TTPKEPAPT 388
            |:  |   | ||| |  | :      |:   ||  : ||  |: |:|    | :|||
Db    692 PEKK-TPVLAKKAPTKPDSEAAADPVSGPSSKDPKLAKKAPVKPRDPSPMKAVPIKPAPK 750

Qy    389 TPKEPAPTTTKKPAPTA-PKEPAPTTPKETAPTTPKKLTPTTP-----EKLAPTTPEKPA 442
          |  |  |   ||| | |  ::|:|    |   | :|  : | ||      :|  |   :  |
Db    751 T--EVPPAVVKKPEPVAKSRDPSPKKAK-AEPNSP--VVPPTPVKNPVKKWKPPWEDDDA 805

Qy    443 PTTPEELAPTTPEEP-----TPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTT-PKE 496
          |  |  :     : |      ||  | :|:|     | |:|  :  | : |:| | : |||
Db    806 PAEPVNVPEPEKKTPVLAKKTPVKPRDPSPKKAVPAKPSTKTDAPPVSVKKPEPVSKPKE 865

Qy    497 PAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPT-TTKEPTSTTSDKPAPTTP 555
          |:|   :  :|  |    || :| |     | |   :  ||   |:|:    :|  |
Db    866 PSPKKAEPNSPVVP----PTPVKNPVKKW-KPPWEDDDEPTEEVKKPSE--PEKKTPVLA 918

Qy    556 KGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPT-------TPKKPAP-----KEL 603
          |   |  ||: ||    :|   :||   |  ||||         :|||  |     :|
Db    919 K-KEPEKPKD-APKVAAKPRDPSPKKAVPE--KEPAKVAAKPRDLSPKKAIPIPANTQEA 974
```

Qy 604 APTTTKGPTS------TTSDKP-----APTTPKET------APTTPKEPAPTTPKKPAPT 646
          || | | |:| || |:| || |::| :||| ||
Db 975 PPTPVKNPVKKWKPPWEDDDEPAEPVSAPEPEKKTPVLAKKAPAKPRDP---SPKKAAPV 1031

Qy 647 TPETPPPTTSEVSTPTTTKEPTTIHKSP----DESTPELSA-EP---TPKALENSPKEPG 698
          : | | || || | | | | || : :|| || || : :| :|
Db 1032 AAK-PDPKIPEV-PPTPVKNPVKKWKPPWEDDDEPSEPVSAPEPEKKTPVLAKKAPTKPA 1089

Qy 699 V-------------PTTKTPAAT------KPEMTTTAKDKTTERDLRTTPETT--TAAPK 737
                      ||:| | : ||: || || : || ||||
Db 1090 TKPDSEAAADPVSGPTSKDPKLSKKAPVEKPKPTTDPKDDKLKPSPAKKPEKAPEPAAPK 1149

Qy 738 MTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTE 797
          | |: | | | || |
Db 1150 KWKPVWDDDPDEPEADFTVPAPSKKPDTEDPADPLG---------GPKTKDPK------- 1193

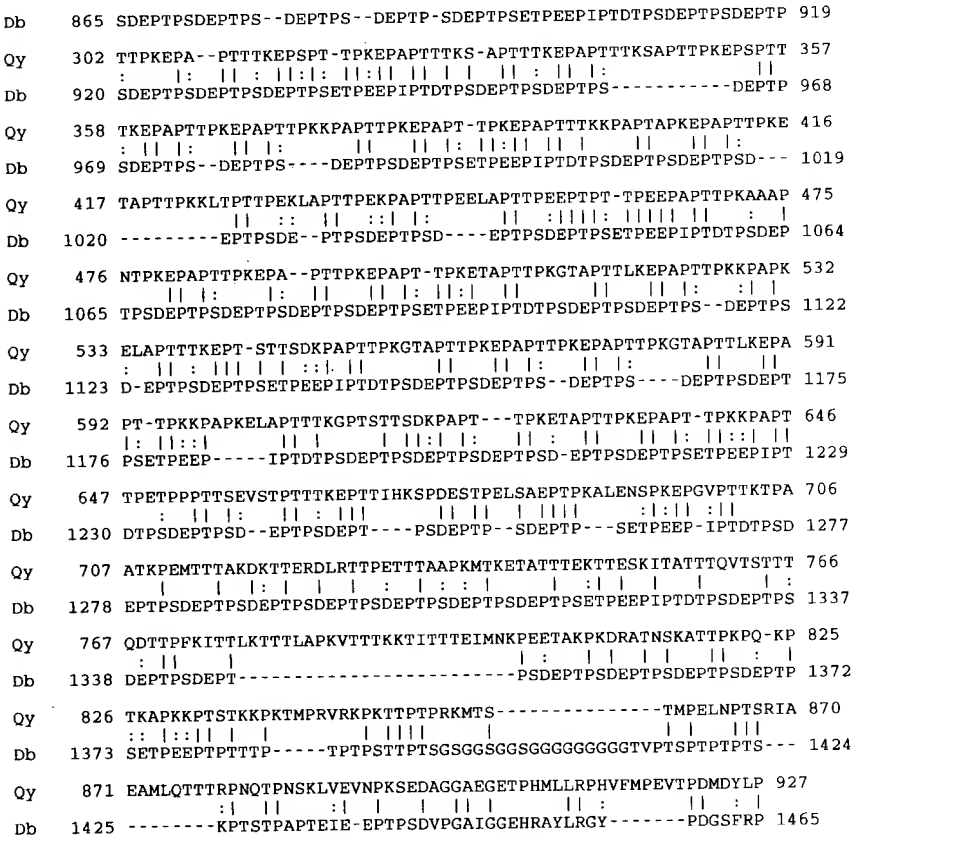
Qy 798 IMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPK--------- 848
          :|| || :: |||:: :| | ||| || |: ||
Db 1194 -LNKKAPAEKPTEK--------PKPKEVSKEPPKPTEPPKP-AAPKKWKPPWEDDPDEPE 1243

Qy 849 ---TTPTPRKMTSTMP 861
           | | |:| : |
Db 1244 ADFTMPAPKKPDTEDP 1259

RESULT 5
S49915
extensin-like protein - maize
C;Species: Zea mays (maize)
C;Date: 05-Mar-1995 #sequence_revision 12-May-1995 #text_change 29-Oct-1999
C;Accession: S49915
R;Rubinstein, A.L.; Broadwater, A.H.; Lowrey, K.; Bedinger, P.A.
submitted to the EMBL Data Library, June 1994
A;Description: Pex genes: pollen-specific genes with extensin-like domains.
A;Reference number: S49915
A;Accession: S49915
A;Status: preliminary
A;Molecule type: DNA
A;Residues: 1-1188 <RUB>
A;Cross-references: EMBL:Z34465; NID:g600117; PIDN:CAA84230.1; PID:g600118

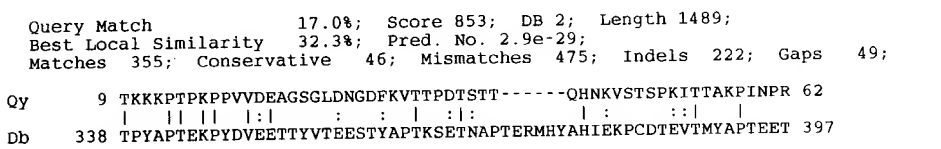
Query Match 15.6%; Score 784; DB 2; Length 1188;
Best Local Similarity 29.9%; Pred. No. 1.9e-26;
Matches 230; Conservative 59; Mismatches 359; Indels 122; Gaps 33;

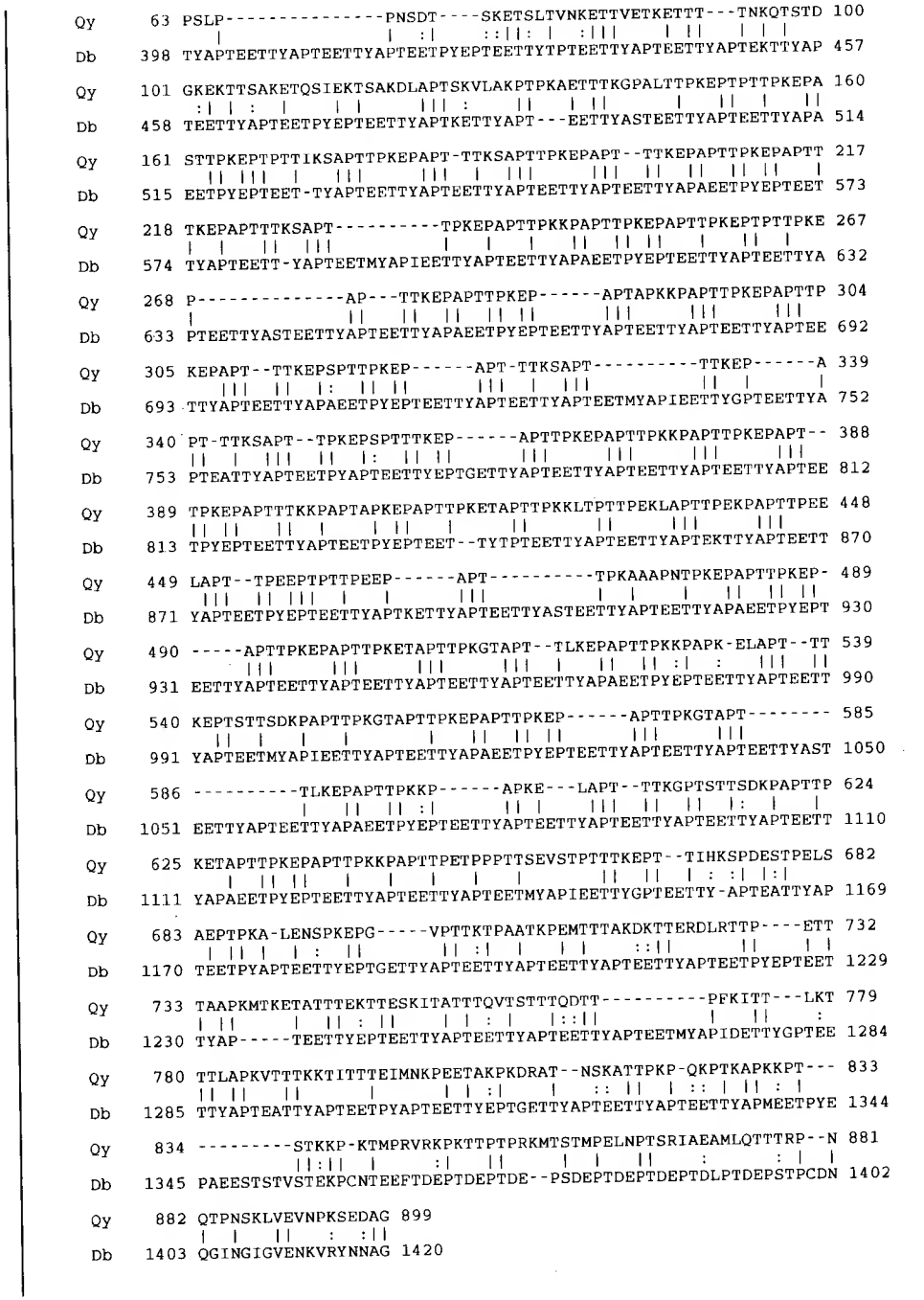
```
Qy     13 PTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPRPSL----PPN 68
          |||  |  |          |:  ||          | ||::       | |||    ||
Db    460 PTPHSPPAD---------DYVPPTPPVPGKSPPATSPSPQVQPPAASTPPPSLVKLSPPQ 510

Qy     69 SDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKV 128
          :                                |     ||||          |     | | |
Db    511 APVG-----------------------SPPPPVKTTSPPAPIG----SPSPPPPVSVV 541

Qy    129 -----LAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPA 183
               :  | | |   :  |   :|  | |      |   :|  | || : | |   | |
Db    542 SPPPPVKSPPPPAPVGSPPPPEKSPPPPAPVASPPPPVKSP--PPPTLVASPPPPVKSPP 599

Qy    184 PTTTKSAPTTP--KEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTP-KEPAPTT 240
          |    ::|  |    | ||    | |  |   :|   | | | |  |:|  | | | |
Db    600 PPAPVASPPPPVKSPPPPTPVASPPPPAPVASSPPPMKSPPPPTPVSSPPPPEKSPPPPP 659

Qy    241 PKKPAPTTPKEPAPTT------PKE---PTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKP 291
          | |  |   : | | |       | |   | ||    | | |   |:|| :| |::|:||
Db    660 PAKSTPPPEEYPTPPTSVKSSPPPEKSLPPPTLIPSPPPQEKPTPPSTPSKP-PSSPEKP 718

Qy    292 APTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPK 351
          :|  ||||  : |: |       | |     | ||   | |      :| : ||:|
Db    719 SP--PKEPVSSPPQTP----KSSPPPAPVSSPPPTPVSSPPALAPVSSPPSVKSSP---- 768

Qy    352 EPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAP 411
            |:| ::  |||     | |     | |  ||   |  |     |   | ||    |
```



RESULT 3
T31108
cyst germination specific acidic repeat protein precursor - Phytophthora infestans
C;Species: Phytophthora infestans (potato late blight agent)
C;Date: 22-Oct-1999 #sequence_revision 22-Oct-1999 #text_change 22-Oct-1999
C;Accession: T31108
R;Goernhardt, B.
submitted to the EMBL Data Library, April 1998
A;Reference number: Z20986
A;Accession: T31108
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-1489 <GOE>
A;Cross-references: EMBL:AF061185; NID:g3851513; PID:g3851514; PIDN:AAC72308.1
C;Genetics:
A;Gene: car90

Query Match 17.0%; Score 853; DB 2; Length 1489;
Best Local Similarity 32.3%; Pred. No. 2.9e-29;
Matches 355; Conservative 46; Mismatches 475; Indels 222; Gaps 49;





J. Clin. Invest. 87, 77-82, 1991
A;Title: Human bronchus and intestine express the same mucin gene.
A;Reference number: A61257; MUID:91086481
A;Accession: A61257
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 'T',1925-1948,'TTS',1952-1954 <JAN>
A;Experimental source: bronchus
R;Xu, G.; Huan, L.; Khatri, I.; Sajjan, U.S.; McCool, D.; Wang, D.; Jones, C.; Forstner,
Biochem. Biophys. Res. Commun. 183, 821-828, 1992
A;Title: Human intestinal mucin-like protein (MLP) is homologous with rat MLP in the C-t
A;Reference number: PQ0328; MUID:92198477
A;Accession: PQ0328
A;Molecule type: mRNA
A;Residues: 2328-2468 <XUG>
A;Cross-references: GB:M86523
A;Experimental source: small intestine
A;Accession: PQ0329
A;Molecule type: protein
A;Residues: 2328-2342,'K',2344-2354 <XUG1>
C;Genetics:
A;Gene: GDB:MUC2
A;Cross-references: GDB:120203; OMIM:158370
A;Map position: 11p15.5-11p15.5
C;Superfamily: von Willebrand factor; von Willebrand factor type A repeat homology; von
C;Keywords: glycoprotein; intestine; tandem repeat
F;2766-2834/Domain: von Willebrand factor type C repeat homology <VWC>

Query Match 21.2%; Score 1062.5; DB 2; Length 3020;
Best Local Similarity 29.2%; Pred. No. 8e-38;
Matches 342; Conservative 74; Mismatches 442; Indels 315; Gaps 39;

```
Qy      32 FKVTTPDTSTTQH---NKVSTSPKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETK 88
           :::  |: :  :|     ::| |   |       |: |        || |    ||  |
Db    1240 WEICGPNGTVEKHFNICSITTRPSTLTTFTTITLPTTP--------TSFTTTTTTTTPTS 1291

Qy      89 ET--TTTNK--------------QTSTDGKEKTTS--AKETQSIEKTSAKD-------- 121
            |  :|| |                | ||  :          : ||  | ||
Db    1292 STVLSTTPKLCCLWSDWINEDHPSSGSDDGDREPFDGVCGAPEDIECRSVKDPHLSLEQH 1351

Qy     122 ---------------------------------------------LAPTSKVLAKPTPKAE 137
                                                         :   |     |:|
Db    1352 GQKVQCDVSVGFICKNEDQFGNGPFGLCYDYKIRVNCCWPMDKCITTPSPPTTTPSPPPT 1411

Qy     138 TTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAP-TTPKEPAPTTTKSAPTTPKE 196
           |||  |  |||  || ||   | :|||  |  ||    | |||  |  |||   |||
Db    1412 TTTTLPPTTTPSPPTTTTTTPPPTTTPSPPITTTTTPLPTTTPSPPISTTTTPPPTTTPS 1471

Qy     197 PAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAP-TTPKEPAPTTPKKPAPTTPKEPAPT 255
           | ||||  |  |||  |  |||  | |||| | | |||  |  :|   |  |||  |  |
Db    1472 P-PTTTPSPPTTTPSPPT-TTTTTPPPTTTPSPPMTTPITPPASTTTLPPTTTPSPPTTT 1529

Qy     256 TPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEP 315
           |   |  |||  |  |   | ||:     ||    | |||   | |||   |  |||  |
Db    1530 TTTPPPTTTPSPPTTTPITP-PTSTTTLPPTTTPSPPPTTTTPPPTTTPSPPTTTTPSP 1588

Qy     316 SPTTPKEPAPTTTKSAP-TTTKEPAPTTTKSAP-TTPKEPSPTTTKEPAPTTPKEPAPTT 373
           |   | |||| | | |||  | |||| | | |||  |  :||  |  |||  |  ||
Db    1589 PTITTTTPPPTTTPSPPTTTTTTPPPTTTPSPPTTTPITPPTSTTTLPPTTTPSPPPTTT 1648

Qy     374 PKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKL 433
           |  |||  |  |||  |  |||  |  | |  |  |||  : |||    | |||
Db    1649 TTPPPTTTPSPPTTTTPSPPITTTTTPPPTTTPSSPITTTP--SPPTT----TMTTPS-- 1700

Qy     434 APTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTT 493
             |||  |  ||    : |||  | |||   |:|||  :    |     |||  |  ||
Db    1701 PTTTPSSPITTTTTPSSTTTP-SPPPTTMTTPSPTTTPSPPTTTMTTLPPTTTSSPLTTT 1759

Qy     494 PKEPAPTTP-----KETAPTTP------------KG------------------------ 512
           |  |: | |       | ||||             |
Db    1760 PLPPSITPPTFSPFSTTTPTTPCVPLCNWTGWLDSGKPNFHKPGGDTELIGDVCGPGWAA 1819
```

```
Qy     513 ---------------------------------------------------------- 512

Db    1820 NISCRATMYPDVPIGQLGQTVVCDVSVGLICKNEDQKPGGVIPMAFCLNYEINVQCCECV 1879

Qy     513 TAPTTLKEPAPTTPKKPAPKELAPTT--TKEPTSTTSDKP----APT-TPKGTAPTTPKE 565
           | |||:     || : | |  :  ||  |  || |::  |    ||| ||  |  |
Db    1880 TQPTTM---TTTTTENPTPTPITTTTTVTPTPTPTSTQSPNGLQAPTPTPISTTTTVTPT 1936

Qy     566 PAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTST-TSDKPAPTTP 624
           | ||  : |  |||  |  |    | ||  : |    :  |||  || | || |    ||
Db    1937 PTPTGTQTPT-TTPITTTTTVTPTPTPTGTQTPTTVLITTTTTMTPTPTPTSTKSTTVTP 1995

Qy     625 KETAPTTPKEPAPTTPKKP-------APTTPETPPPTTSEVSTPTTTK-----EPTTIHK 672
             |  |    | ||  : |         |   || |||     || |      | || :
Db    1996 ITTTTTVTATPTPTGTQTPTMIPISTTTTVTPTPTPTTGSTGPPTHTSTAPIAELTTSNP 2055

Qy     673 SPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETT 732
           |: |||: |   :    |::     :|       :| |  | ||   |:     : | :|
Db    2056 PPESSTPQTSRSTSSPLTESTTLLSTLPPAIEMTSTAPPSTPTAPTTTSGGHTLSPPPST 2115

Qy     733 TAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITT---LKTTTLAPKVTTT 789
           | :|  |    |||   : |  | :| | |:|:    ||  ::|   ::|| | |  ::
Db    2116 TTSPPGTPTRGTTT--GSSSAPTPSTVQTTTTSAWTPTPTPLSTPSIIRTTGLRPYPSSV 2173

Qy     790 KKTITTTEIMNKPEE----------------------------TAKPKDRATNSKATTP 820
                 :   | |                                : |    | || :||
Db    2174 LICCVLNDTYYAPGEEVYNGTYGDTCYFVNCSLSCTLEFYNWSCPSTPSPTPTPSK-STP 2232

Qy     821 KPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRP 880
           | ||:  | |||   ||   |    |:              : | |  : :  : | :
Db    2233 TPSKPSSTPSKPTPGTKPPECPDFDPPR------------QENETWWLCDCFM-ATCKY 2278

Qy     881 NQTPNSKLVEVNP-------------KSEDAGG-----------AEGETPHMLL----- 910
           | |      ||  |             : ||  |               |: || :
Db    2279 NNTVEIVKVECEPPPMPTCSNGLQPVRVEDPDGCCWHWECDCYCTGWGD-PHYVTFDGLY 2337

Qy     911 ------RPHVFMPEVTPDMDYLPRVPNQGIIIN 937
                 :| : |::| :|       | |: |:
Db    2338 YSYQGNCTYVLVEEISPSVD------NFGVYID 2364
```

RESULT 2
T18262
S-layer protein - Clostridium thermocellum
C;Species: Clostridium thermocellum
C;Date: 15-Oct-1999 #sequence_revision 15-Oct-1999 #text_change 15-Oct-1999
C;Accession: T18262
R;Fujino, T.; Beguin, P.; Aubert, J.P.
J. Bacteriol. 175, 1891-1899, 1993
A;Title: Organization of a Clostridium thermocellum gene cluster encoding the cellulo
e.
A;Reference number: Z18847; MUID:93209931
A;Accession: T18262
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-1664 <FUJ>
A;Cross-references: EMBL:X67506; NID:g296879; PID:g296881; PIDN:CAA47841.1

Query Match 18.8%; Score 942; DB 2; Length 1664;
Best Local Similarity 33.0%; Pred. No. 5.6e-33;
Matches 276; Conservative 83; Mismatches 312; Indels 166; Gaps 43;

```
Qy     128 VLAKPTP-KAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTT 186
           |: :| | || :    |  |  ||||:          ||||:     ||   || |:
Db     758 VVIQPAPIKAASDEPIPTDTPSDEPTPS----------DEPTPS---DEPTPSDEPTPSD 804

Qy     187 TKSAPTTPKEPAPTTTKEPAPTTPKEPA----PTTTKEPAPTTTKSAPTTPKEPAPTTPK 242
             :  ||:|| || |    ||   ||     || : || |:   :   ||:|| ||
Db     805 EPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTP 864

Qy     243 KPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPT-TPKEPAPTAPKKPAPTTPKEPAP 301
             ||   || |:   ||||:   || | : || |: ||:|| ||      ||   || |
```

GenCore version 4.5
Copyright (c) 1993 - 2000 Compugen Ltd.

OM protein - protein search, using sw model

Run on: April 26, 2002, 16:13:30 ; Search time 82.85 Seconds
(without alignments)
865.181 Million cell updates/sec

Title: US-09-556-246-1_COPY_200_1140
Perfect score: 5011
Sequence: 1 VKDNKKNRTKKKPTPKPPVV.....DMDYLPRVPNQGIIINPMLS 941

Scoring table: BLOSUM62
Gapop 10.0 , Gapext 0.5

Searched: 219241 seqs, 76174552 residues

Total number of hits satisfying chosen parameters: 219241

Minimum DB seq length: 0
Maximum DB seq length: 2000000000

Post-processing: Minimum Match 0%
Maximum Match 100%
Listing first 45 summaries

Database : PIR_68:*
1: pir1:*
2: pir2:*
3: pir3:*
4: pir4:*

Pred. No. is the number of results predicted by chance to have a score greater than or equal to the score of the result being printed, and is derived by analysis of the total score distribution.

SUMMARIES

Result No.	Score	% Query Match	Length	DB	ID	Description
1	1062.5	21.2	3020	2	A43932	mucin 2 precursor,
2	942	18.8	1664	2	T18262	S-layer protein -
3	853	17.0	1489	2	T31108	cyst germination s
4	802	16.0	1274	2	T16251	hypothetical prote
5	784	15.6	1188	2	S49915	extensin-like prot
6	776	15.5	2187	2	T30826	nascent polypeptid
7	762.5	15.2	1367	1	S48478	glucan 1,4-alpha-g
8	671	13.4	1151	2	T18535	high molecular mas
9	666.5	13.3	1344	1	A35175	mucin 1 precursor,
10	659.5	13.2	1229	2	T25697	hypothetical prote
11	633	12.6	489	2	T11622	extensin class 1 p
12	632	12.6	761	2	C84672	hypothetical prote
13	632	12.6	7962	2	I38346	elastic titin - hu
14	629	12.6	3507	2	T34513	hypothetical prote
15	626.5	12.5	990	2	I51618	nucleolar phosphop
16	622.5	12.4	839	2	F75518	hypothetical prote
17	622	12.4	6642	2	T29757	protein UNC-89 - C
18	620.5	12.4	971	2	T19431	hypothetical prote
19	617	12.3	3570	2	T45025	mucin MUC5B, trach
20	607.5	12.1	801	2	T29018	hypothetical prote
21	607.5	12.1	924	2	S27923	gene LF3 protein -
22	605	12.1	379	2	S50125	larval glue protei
23	559.5	11.2	350	2	S22456	hydroxyproline-ric
24	556.5	11.1	856	2	T16543	hypothetical prote
25	556.5	11.1	875	2	S23760	polyphenolic adhes
26	555.5	11.1	1630	2	A53577	ascites sialoglyco
27	551	11.0	620	2	S06733	hydroxyproline-ric
28	549	11.0	2232	2	T34434	hypothetical prote
29	545	10.9	873	2	A47283	calphotin - fruit
30	542	10.8	369	2	S20500	hydroxyproline-ric
31	538.5	10.7	416	2	JU0465	extensin precursor
32	533	10.6	1087	1	QFMSH	neurofilament trip
33	532	10.6	756	2	T27642	hypothetical prote
34	530.5	10.6	1162	2	JH0557	exo-alpha-sialidas
35	530	10.6	865	2	A47282	calcium-binding pr
36	522	10.4	328	2	JQ0985	hydroxyproline-ric
37	518	10.3	813	2	S70795	vsaA protein precu
38	518	10.3	1072	1	A37221	neurofilament trip
39	512	10.2	866	2	T45462	membrane glycoprot
40	509.5	10.2	662	2	A45155	mucin FIM-C.1 - Af
41	509.5	10.2	867	2	T45463	membrane glycoprot
42	509	10.2	1459	2	T32271	hypothetical prote
43	505	10.1	700	2	A54641	interspersed repea
44	504.5	10.1	606	2	A43427	neurofilament trip
45	503.5	10.0	1118	2	A48292	mucin, tracheobron

ALIGNMENTS

RESULT 1
A43932
mucin 2 precursor, intestinal - human (fragments)
N;Alternate names: mucin SMUC-41
C;Species: Homo sapiens (man)
C;Date: 10-Mar-1993 #sequence_revision 12-Apr-1996 #text_change 05-Nov-1999
C;Accession: A49963; A45106; B45106; A43932; B33532; A61257; PQ0328; PQ0329
R;Gum Jr., J.R.; Hicks, J.W.; Toribara, N.W.; Siddiki, B.; Kim, Y.S.
J. Biol. Chem. 269, 2440-2446, 1994
A;Title: Molecular cloning of human intestinal mucin (MUC2) cDNA. Identification of t
A;Reference number: A49963; MUID:94132002
A;Accession: A49963
A;Molecule type: mRNA
A;Residues: 1-639 <GU1>
A;Cross-references: GB:L21998
R;Gum Jr., J.R.; Hicks, J.W.; Toribara, N.W.; Rothe, E.M.; Lagace, R.E.; Kim, Y.S.
J. Biol. Chem. 267, 21375-21383, 1992
A;Title: The human MUC2 intestinal mucin has cysteine-rich subdomains located both up
A;Reference number: A45106; MUID:93016075
A;Accession: A45106
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 626-1895 <GU2>
A;Cross-references: GB:M94131; NID:g186395; PIDN:AAA59163.1; PID:g186396
A;Note: sequence extracted from NCBI backbone (NCBIP:116706)
A;Accession: B45106
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 2037-3020 <GU3>
A;Cross-references: GB:M94132; NID:g186397; PIDN:AAA59164.1; PID:g186398
A;Experimental source: colon
A;Note: sequence extracted from NCBI backbone (NCBIP:116698)
R;Toribara, N.W.; Gum Jr., J.R.; Culhane, P.J.; Lagace, R.E.; Hicks, J.W.; Petersen,
J. Clin. Invest. 88, 1005-1013, 1991
A;Title: MUC-2 human small intestinal mucin gene structure. Repeated arrays and polym
A;Reference number: A43932; MUID:91358717
A;Accession: A43932
A;Molecule type: DNA
A;Residues: 1343-1350,'L',1352-1411,'S',1413-1448,'P',1450-1503,'T',1505-1915 <TOR>
A;Cross-references: GB:M74027; NID:g188863; PIDN:AAA59875.1; PID:g188864
A;Note: sequence inconsistent with the nucleotide translation
A;Note: sequence extracted from NCBI backbone (NCBIN:55749, NCBIP:55750)
R;Gum, J.R.; Byrd, J.C.; Hicks, J.W.; Toribara, N.W.; Lamport, D.T.A.; Kim, Y.S.
J. Biol. Chem. 264, 6480-6487, 1989
A;Title: Molecular cloning of human intestinal mucin cDNAs. Sequence analysis and evi
A;Reference number: A33532; MUID:89197956
A;Accession: B33532
A;Molecule type: mRNA
A;Residues: 1916-2193 <GU4>
A;Cross-references: GB:M22405; NID:g188873; PIDN:AAA36334.1; PID:g188874
A;Experimental source: intestine
R;Jany, B.H.; Gallup, M.W.; Yan, P.S.; Gum, J.R.; Kim, Y.S.; Basbaum, C.B.

```
Db    472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy    589 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 636
Db    531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy    637 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 682
Db    586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy    683 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 726
Db    645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy    727 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 772
Db    703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy    773 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 828
Db    763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy    829 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 888
Db    821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy    889 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQT 948
Db    869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPX------------PGST 911

Qy    949 -PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSD 1007
Db    912 APXAHGVTSAPDXRPXPGSTAPXAH----------GVTSAPDXRPXPGSTA----PXAHG 957

Qy   1008 ETNICNGKPVDGLTT-LRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGIPSPID 1062
Db    958 VTSAPDXRPXPGSTAPXAHGVTSA----------PDXRPXPGSTAPXAHGVTSAPD 1003


RESULT 15
US-08-276-967-2
; Sequence 2, Application US/08276967
; Patent No. 5851817
;  GENERAL INFORMATION:
;    APPLICANT:  Hardy, Daniel M.
;    APPLICANT:  Garbers, David L.
;    TITLE OF INVENTION:  Species-Specific Egg-Binding Proteins of
;    TITLE OF INVENTION:  Sperm
;    NUMBER OF SEQUENCES:  8
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  Arnold, White & Durkee
;      STREET:  P. O. Box 4433
;      CITY:  Houston
;      STATE:  Texas
;      COUNTRY:  USA
;      ZIP:  77210-4433
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.25
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/08/276,967
;      FILING DATE:  Submitted Herewith
;      CLASSIFICATION:  530
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Kitchell, Barbara S.
;      REGISTRATION NUMBER:  33,928
;      REFERENCE/DOCKET NUMBER:  UTSD:418\KIT
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  713-787-1400
;      TELEFAX:  713-789-2679
;      TELEX:  79-0924
;  INFORMATION FOR SEQ ID NO:  2:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  2476 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
US-08-276-967-2


  Query Match             6.2%;  Score 419.5;  DB 2;  Length 2476;
  Best Local Similarity   32.5%;  Pred. No. 1.4e-19;
  Matches  163;  Conservative   43;  Mismatches  177;  Indels  119;  Gaps   31;

Qy    471 PKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEP 530
Db    313 PSETSVSTEKPVAPTE----KPTVPSEIYTIPTEKPMVHMEKPIVHT--EKPTVPT-EKP 365

Qy    531 APTTPKAAAPNTPKEPA----PTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEP 586
Db    366 TIPTEKSTVPT--KKPTVFKEPTLPPE-GPTVPAE-RPTTPPEGPAVPPKG--PTVLTE- 418

Qy    587 APTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPT-TPKEPAPTTPKEPAPTTPK 645
Db    419 ------------WPTSHTEKSTVHTEKPILPTGKSTIPTEKPMVPTKRT------TTP- 458

Qy    646 GTAPTTLKEPAPTTPKKPAPKELAPT-TTKGPTSTT---SDKPAPTTPKETAPTTPKEPA 701
Db    459 -TERTTIPAEKPTV---PIEKPMVPTERTTIPTERTTIPTEKPTVPTEKLTVPT--EKPI 512

Qy    702 PTTPKKPAPTTPETPPP-----TTSEVSTPT-----TTKEPTTIHKSPDESTPELSAEPT 751
Db    513 VPTEKPIVPTEKHTIPTEKLTVLTERTTTPTERTTIPTEKPTVPTEKPSVPT-EKPTVPT 571

Qy    752 PKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTT 811
Db    572 --------EEPTIPTEKLTVPT--ERTTTPTKRTTTPTIRTTTPTI--------RTTTPT 613

Qy    812 EKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETA 871
Db    614 ERTT------TPTIRTTTPTERTT---IPTKKTT------VPTEKTIIPT------ERTI 652

Qy    872 KPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPT 931
Db    653 AP---------TTPQP-SPTLVPTQPAAVVMPST------SATTVTPRTTIASCP---PN 693

Qy    932 SRIAEAMLQTTTRPNQTPNSKL 953
Db    694 AHFERCACPVSCQ-SPTPNCEL 714


Search completed: April 26, 2002, 16:25:49
Job time: 489 sec
```

```
Qy    773 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 828
Db    763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy    829 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 888
Db    821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy    889 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQT 948
Db    869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPX------------PGST 911

Qy    949 -PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSD 1007
Db    912 APXAHGVTSAPDXRPXPGSTAPXAH----------GVTSAPDXRPXPGSTA----PXAHG 957

Qy   1008 ETNICNGKPVDGLTT-LRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGIPSPID 1062
Db    958 VTSAPDXRPXPGSTAPXAHGVTSA----------PDXRPXPGSTAPXAHGVTSAPD 1003
```

```
RESULT  14
US-09-083-116-2
; Sequence 2, Application US/09083116
; Patent No. 6203795
;  GENERAL INFORMATION:
;    APPLICANT:  CHAMBON, Pierre
;    APPLICANT:  KIENY, Marie-Paule
;    APPLICANT:  LATHE, Richard
;    APPLICANT:  HAREUVENI, Mara
;    TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;    TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
;    NUMBER OF SEQUENCES:  5
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
;      STREET:  P.O. Box 1404
;      CITY:  Alexandria
;      STATE:  Virginia
;      COUNTRY:  United States
;      ZIP:  22313-1404
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/09/083,116
;      FILING DATE:
;      CLASSIFICATION:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  08/479,537
;      FILING DATE:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  WO PCT/FR91/00835
;      FILING DATE:  23-OCT-1991
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/039,320
;      FILING DATE:  04-APR-1993
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/403,576
;      FILING DATE:  14-MAR-1995
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Teskin, Robin L.
;      REGISTRATION NUMBER:  35,030
;      REFERENCE/DOCKET NUMBER:  017753-025
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (703) 836-6620
;      TELEFAX:  (703) 836-2021
;  INFORMATION FOR SEQ ID NO:  2:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  2035 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  peptide
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  128..1899
;      OTHER INFORMATION:  /note= "The amino acids spanning
;      OTHER INFORMATION:  128 to 1899 constitute a repeated region wherein the repea
;      OTHER INFORMATION:  20 amino acids, 17 of which are fixed. The number of such
;      OTHER INFORMATION:  repeats varies from 1 to 40."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  134
;      OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;      OTHER INFORMATION:  Xaa Xaa which is the codon for Pro or Ala wherein Pro = CC
;      OTHER INFORMATION:  CCC, CCA, or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  144
;      OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;      OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC,
;      OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  147
;      OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC,
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  1..21
;      OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;      OTHER INFORMATION:  21 amino acid precursor sequence."
US-09-083-116-2
```

```
  Query Match             6.2%;  Score 424.5;  DB 4;  Length 2035;
  Best Local Similarity   23.7%;  Pred. No. 5.3e-20;
  Matches  255;  Conservative   69;  Mismatches  511;  Indels  241;  Gaps   53;

Qy    140 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 198
Db     16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy    199 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 241
Db     66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy    242 TTPKEPAPTTT-------KSAPTTPKEP-----------APTTTKEPAPTTPKEPAPTT 282
Db    123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy    283 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 329
Db    181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy    330 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 376
Db    241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy    377 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 425
Db    301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy    426 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 477
Db    360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy    478 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 533
Db    420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy    534 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 588
```

```
Db    958 VTSAPDXRPXPGSTAPXAHGVTSA----------PDXRPXPGSTAPXAHGVTSAPD 1003

RESULT 13
US-08-479-537A-2
; Sequence 2, Application US/08479537A
; Patent No. 5861381
; GENERAL INFORMATION:
;   APPLICANT:  CHAMBON, Pierre
;   APPLICANT:  KIENY, Marie-Paule
;   APPLICANT:  LATHE, Richard
;   APPLICANT:  HAREUVENI, Mara
;   TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;   TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
;   NUMBER OF SEQUENCES:  5
;   CORRESPONDENCE ADDRESS:
;     ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
;     STREET:  P.O. Box 1404
;     CITY:  Alexandria
;     STATE:  Virginia
;     COUNTRY:  United States
;     ZIP:  22313-1404
;   COMPUTER READABLE FORM:
;     MEDIUM TYPE:  Floppy disk
;     COMPUTER:  IBM PC compatible
;     OPERATING SYSTEM:  PC-DOS/MS-DOS
;     SOFTWARE:  PatentIn Release #1.0, Version #1.30
;   CURRENT APPLICATION DATA:
;     APPLICATION NUMBER:  US/08/479,537A
;     FILING DATE:  07-JUN-1995
;     CLASSIFICATION:  514
;   PRIOR APPLICATION DATA:
;     APPLICATION NUMBER:  FR 90/13101
;     FILING DATE:  23-OCT-1990
;   PRIOR APPLICATION DATA:
;     APPLICATION NUMBER:  WO PCT/FR91/00835
;     FILING DATE:  23-OCT-1991
;   PRIOR APPLICATION DATA:
;     APPLICATION NUMBER:  US 08/039,320
;     FILING DATE:  04-APR-1993
;   PRIOR APPLICATION DATA:
;     APPLICATION NUMBER:  US 08/403,576
;     FILING DATE:  14-MAR-1995
;   ATTORNEY/AGENT INFORMATION:
;     NAME:  Teskin, Robin L.
;     REGISTRATION NUMBER:  35,030
;     REFERENCE/DOCKET NUMBER:  017753-025
;   TELECOMMUNICATION INFORMATION:
;     TELEPHONE:  (703) 836-6620
;     TELEFAX:  (703) 836-2021
; INFORMATION FOR SEQ ID NO:  2:
;   SEQUENCE CHARACTERISTICS:
;     LENGTH:  2035 amino acids
;     TYPE:  amino acid
;     STRANDEDNESS:  single
;     TOPOLOGY:  linear
;   MOLECULE TYPE:  peptide
;   FEATURE:
;     NAME/KEY:  Peptide
;     LOCATION:  128..1899
;     OTHER INFORMATION:  /note= "The amino acids spanning
;     OTHER INFORMATION:  128 to 1899 constitute a repeated region wherein the repeat i
;     OTHER INFORMATION:  20 amino acids, 17 of which are fixed. The number of such
;     OTHER INFORMATION:  repeats varies from 1 to 40."
;   FEATURE:
;     NAME/KEY:  Peptide
;     LOCATION:  134
;     OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;     OTHER INFORMATION:  Xaa Xaa which is the codon for Pro or Ala wherein Pro = CCT,
;     OTHER INFORMATION:  CCC, CCA, or CCG; and Ala = GCT, GCC, GCA, or GCG."
;   FEATURE:
;     NAME/KEY:  Peptide
;     LOCATION:  144
;     OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;     OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC,
;     OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;   FEATURE:
;     NAME/KEY:  Peptide
;     LOCATION:  147
;     OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;     OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC,
;     OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;   FEATURE:
;     NAME/KEY:  Peptide
;     LOCATION:  1..21
;     OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;     OTHER INFORMATION:  21 amino acid precursor sequence."
US-08-479-537A-2

  Query Match             6.2%;  Score 424.5;  DB 2;  Length 2035;
  Best Local Similarity   23.7%;  Pred. No. 5.3e-20;
  Matches  255;  Conservative   69;  Mismatches  511;  Indels  241;  Gaps   53;

Qy    140 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 198
Db     16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy    199 P-KAETTTKGP--ALTTPKEP--------------TPTTPKEPASTTPKEPTPTTIKSAP 241
Db     66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy    242 TTPKEPAPTTT-------KSAPTTPKEP------------APTTTKEPAPTTPKEPAPTT 282
Db    123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy    283 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 329
Db    181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy    330 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 376
Db    241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy    377 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 425
Db    301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy    426 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 477
Db    360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy    478 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 533
Db    420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy    534 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 588
Db    472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy    589 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 636
Db    531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy    637 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 682
Db    586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy    683 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 726
Db    645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy    727 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 772
Db    703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762
```

```
;       STREET:  P.O. Box 1404
;       CITY:  Alexandria
;       STATE:  Virginia
;       COUNTRY:  United States
;       ZIP:  22313-1404
;     COMPUTER READABLE FORM:
;       MEDIUM TYPE:  Floppy disk
;       COMPUTER:  IBM PC compatible
;       OPERATING SYSTEM:  PC-DOS/MS-DOS
;       SOFTWARE:  PatentIn Release #1.0, Version #1.30
;     CURRENT APPLICATION DATA:
;       APPLICATION NUMBER:  US/09/083,116
;       FILING DATE:
;       CLASSIFICATION:
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  08/479,537
;       FILING DATE:
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  WO PCT/FR91/00835
;       FILING DATE:  23-OCT-1991
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  US 08/039,320
;       FILING DATE:  04-APR-1993
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  US 08/403,576
;       FILING DATE:  14-MAR-1995
;     ATTORNEY/AGENT INFORMATION:
;       NAME:  Teskin, Robin L.
;       REGISTRATION NUMBER:  35,030
;       REFERENCE/DOCKET NUMBER:  017753-025
;     TELECOMMUNICATION INFORMATION:
;       TELEPHONE:  (703) 836-6620
;       TELEFAX:  (703) 836-2021
;   INFORMATION FOR SEQ ID NO:  5:
;     SEQUENCE CHARACTERISTICS:
;       LENGTH:  1867 amino acids
;       TYPE:  amino acid
;       STRANDEDNESS:  single
;       TOPOLOGY:  linear
;     MOLECULE TYPE:  peptide
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  128..1727
;       OTHER INFORMATION:  /note= "The amino acids spanning
;       OTHER INFORMATION:  128 to 1727 constitute a repeated region wherein the repeat i
;       OTHER INFORMATION:  20 amino acids, 17 of which are fixed.  The number of such
;       OTHER INFORMATION:  repeats varies from 1 to 40."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  134
;       OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;       OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;       OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  144
;       OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;       OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC, ACA
;       OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  147
;       OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;       OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;       OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  1..21
;       OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;       OTHER INFORMATION:  21 amino acid precursor sequence."
US-09-083-116-5
```

```
  Query Match             6.2%;  Score 424.5;  DB 4;  Length 1867;
  Best Local Similarity   23.7%;  Pred. No. 4.8e-20;
  Matches  255;  Conservative   69;  Mismatches  511;  Indels  241;  Gaps   53;

Qy    140 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 198
          | |||          | :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db     16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy    199 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 241
          |    :||:|    |   ||                | |     ||||     |   |||
Db     66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy    242 TTPKEPAPTTT-------KSAPTTPKEP------------APTTTKEPAPTTPKEPAPTT 282
             :||| :|         |||      |              ||      | | |     |:
Db    123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy    283 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 329
            :  | | :|         |||      | | |         ||      | | |     |:
Db    181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy    330 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 376
            :  | | :  |         ||      | |||         ||      | | |     |:
Db    241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy    377 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 425
            :  | |   ||       |||      | |         |||      | |
Db    301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy    426 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 477
            ||      | | |      ||      | | |         ||      | |||      |
Db    360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy    478 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 533
            :  :  |    | |      |:    ||      | | |      |:  :  |     | |
Db    420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy    534 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 588
            | |     :||:    |  | |      |:  :  |  |  |||       | ||     |
Db    472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy    589 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 636
            | |           :| |   ||       ||     :| |  | |     |:    ||
Db    531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy    637 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 682
            |  | |       :||      |  | |           :| |   ||     | ||
Db    586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy    683 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 726
            :| |  | |       :||      | | |           :| | :  | |     | |
Db    645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy    727 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 772
            | :       :  :||      | | |     : ::|      | :  | |
Db    703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy    773 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 828
             :|   :||       |:  | |  | :|  ||      |:      :           |||
Db    763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy    829 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 888
                |     |   | ||       :  :     |  ||       |::    | |
Db    821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy    889 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQT 948
             : ||     |  |   |      | |    :||  |: |               | |
Db    869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPX------------PGST 911

Qy    949 -PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSD 1007
             | :  |    |      |:     |              ||   | |  :      |
Db    912 APXAHGVTSAPDXRPXPGSTAPXAH----------GVTSAPDXRPXPGSTA----PXAHG 957

Qy   1008 ETNICNGKPVDGLTT-LRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGIPSPID 1062
```

```
;       APPLICATION NUMBER:  WO PCT/FR91/00835
;       FILING DATE:  23-OCT-1991
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  US 08/039,320
;       FILING DATE:  04-APR-1993
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  US 08/403,576
;       FILING DATE:  14-MAR-1995
;     ATTORNEY/AGENT INFORMATION:
;       NAME:  Teskin, Robin L.
;       REGISTRATION NUMBER:  35,030
;       REFERENCE/DOCKET NUMBER:  017753-025
;     TELECOMMUNICATION INFORMATION:
;       TELEPHONE:  (703) 836-6620
;       TELEFAX:  (703) 836-2021
;   INFORMATION FOR SEQ ID NO:  5:
;     SEQUENCE CHARACTERISTICS:
;       LENGTH:  1867 amino acids
;       TYPE:  amino acid
;       STRANDEDNESS:  single
;       TOPOLOGY:  linear
;     MOLECULE TYPE:  peptide
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  128..1727
;       OTHER INFORMATION:  /note= "The amino acids spanning
;       OTHER INFORMATION:  128 to 1727 constitute a repeated region wherein the repeat i
;       OTHER INFORMATION:  20 amino acids, 17 of which are fixed.  The number of such
;       OTHER INFORMATION:  repeats varies from 1 to 40."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  134
;       OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;       OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;       OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  144
;       OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;       OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC, ACA
;       OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  147
;       OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;       OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;       OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  1..21
;       OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;       OTHER INFORMATION:  21 amino acid precursor sequence."
US-08-479-537A-5


  Query Match             6.2%;  Score 424.5;  DB 2;  Length 1867;
  Best Local Similarity   23.7%;  Pred. No. 4.8e-20;
  Matches  255;  Conservative   69;  Mismatches  511;  Indels  241;  Gaps   53;

Qy    140 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 198
          | |||           | :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db     16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy    199 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 241
          |    :||:|    |    ||                | |     ||||     |   |||
Db     66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy    242 TTPKEPAPTTT-------KSAPTTPKEP------------APTTTKEPAPTTPKEPAPTT 282
            :||| :|         |||     |              ||      | | |     |:
Db    123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy    283 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 329
            :  | | :|        |||     |  | | |       ||      | | |     |:

Db    181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy    330 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 376
            :  | | :  |        ||      |  |||      ||      |  | |    |:
Db    241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy    377 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 425
            :  | |  ||       |||      | |         |||      |  |
Db    301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy    426 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 477
          ||      |  | |       ||      | | |       ||      |  |||     |
Db    360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy    478 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 533
          :  :  |    | |       |:     ||      |  | |      |: :  |     | |
Db    420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy    534 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 588
          | |     :||:     |  | |      |: :  |  |  |||      | ||      |
Db    472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy    589 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 636
          | |           :| |   ||      ||    :| |  | |     |:   ||
Db    531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy    637 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 682
          |  | |       :||     |  | |              :| |   ||     | ||
Db    586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy    683 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 726
          :| |  | |       :||      |  | |            :| | :  | |      | |
Db    645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy    727 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 772
              | :       :  :||     |  |    : ::|     | :  | |
Db    703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy    773 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 828
           :|   :||       |:  | |  | :|  ||     |:   :                |||
Db    763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy    829 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 888
                |       |  | ||      :  :    |  ||      |::    | |
Db    821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy    889 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQT 948
            : ||      |  |  |       | |    :||  |:  |                | |
Db    869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPX------------PGST 911

Qy    949 -PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSD 1007
           | :  |   |       |:     |              ||   |  |  :      |
Db    912 APXAHGVTSAPDXRPXPGSTAPXAH----------GVTSAPDXRPXPGSTA----PXAHG 957

Qy   1008 ETNICNGKPVDGLTT-LRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGIPSPID 1062
          |:  : :|  | |    :|   |          |   | |          |: | |
Db    958 VTSAPDXRPXPGSTAPXAHGVTSA----------PDXRPXPGSTAPXAHGVTSAPD 1003


RESULT  12
US-09-083-116-5
; Sequence 5, Application US/09083116
; Patent No. 6203795
;  GENERAL INFORMATION:
;    APPLICANT:  CHAMBON, Pierre
;    APPLICANT:  KIENY, Marie-Paule
;    APPLICANT:  LATHE, Richard
;    APPLICANT:  HAREUVENI, Mara
;    TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;    TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
;    NUMBER OF SEQUENCES:  5
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
;
```

```
Qy    442 PA--PTTPKEPA--PTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEK 497
          |:  ||   :|:  ||   :|:   | |  |: |    ||   : :     |: ||  |
Db    309 PSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYP----PTYKAKPSYPPTYKVKPT--YK 362

Qy    498 LAPTTPE--KPAPTTPEELAPTTPEEPT--PTTPEEPA-PTTPKAAAPNTPKEPAPTTPK 552
            || |   |  |: |    ||   :|:  ||   :|: | | | | |: |    ||
Db    363 AKPTYPSTYKAKPSYP----PTYKAKPSYPPTYKAKPSYPPTYK-AKPSYP----PTYKA 413

Qy    553 EPA-PTT----PKEPAPTTPKETAPTT--PKGTAPTTLKEPAPTTPKKPAPKELAPTTTK 605
          :|  |:|    |  |    || : | |   | : |:| |  :   |   |     ||
Db    414 KPTYPSTYKAKPSYPPTYKPKISYPPTYKAKPSYPSTYKAKSSYPPTYKAKPSYPPTYKA 473

Qy    606 EPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPA--PTTPKKP 663
          :||  :: |  ||      |: |    ||   :|:        ||   :|:  ||   ||
Db    474 KPTYPSTYKAKPTYK--AKPSYP----PTYKAKPS------YPPTYKAKPSYPPTYKAKP 521

Qy    664 APKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPA--PTTPKKPA-PTT----PETP 716
            |   ||    |:   : |  |: |    ||   :|:  ||   ||  |:|    |  |
Db    522 TYK-AKPTYKAKPSYPPTYKAKPSYP----PTYKAKPSYPPTYKAKPTYPSTYKAKPSYP 576

Qy    717 PPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPE 776
          |   :: | | |      :|:     |   |:|| ||      :|  |:|      ||
Db    577 PTYKAKPSYPPT-------YKAKPSYPPTYKAKPTYKA------KPTYPST---YKAKPS 620

Qy    777 MTTTAKDKTTERDLRTTPETTTAAP 801
            | | | |  :   : | |  | |
Db    621 YPPTHKAKPTYKAKPSYPPTHKAKP 645


RESULT  10
US-09-219-849-5
; Sequence 5, Application US/09219849
; Patent No. 6150081
; GENERAL INFORMATION:
;  APPLICANT:  VAN HEERDE, GEORGE V.
;  APPLICANT:  VAN RIJN, ALEXIS C.
;  APPLICANT:  BOUWSTRA, JAN B.
;  APPLICANT:  DE WOLF, FREDERIK A.
;  APPLICANT:  MOOBROEK, ANDREAS
;  APPLICANT:  WERTEN, MARC W.T.
;  APPLICANT:  WIND, RICHELE D.
;  APPLICANT:  VAN DEN BOSCH, TANJA J.
;  TITLE OF INVENTION:  SILVER HALIDE EMULSIONS WITH RECOMBINANT COLLAGEN
;  TITLE OF INVENTION:  SUITABLE FOR PHOTOGRAPHIC APPLICATION AND ALSO THE
;  TITLE OF INVENTION:  PREPARATION THEREOF
;  FILE REFERENCE:  2728-2
;  CURRENT APPLICATION NUMBER: US/09/219,849
;  CURRENT FILING DATE:  1998-12-23
;  NUMBER OF SEQ ID NOS:  50
;  SOFTWARE:  PatentIn Ver. 2.1
; SEQ ID NO 5
;   LENGTH: 960
;   TYPE: PRT
;   ORGANISM: Artificial Sequence
;   FEATURE:
;   OTHER INFORMATION: Description of Artificial Sequence: Illustrative
;   OTHER INFORMATION: amino acid sequence
US-09-219-849-5


  Query Match             6.6%;  Score 452;  DB 4;  Length 960;
  Best Local Similarity   25.6%;  Pred. No. 3.4e-22;
  Matches  173;  Conservative   60;  Mismatches  309;  Indels  134;  Gaps   41;

Qy    197 PTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAP--TTTKS 254
          | |     ::|    |  | |  |  | |  |  |       : |   ::| |      :
Db     72 PGPAGPPGSRDPG--PPGAPGPAGP--PGSRDPGPPGAPG-PAGPPGSRDPGPPGAPGPA 126

Qy    255 APTTPKEPAPTTTKEPA-PTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTP---KKPA 310
          |   ::| |     || |   ::| |      ||        |  |  | |  |   : |
Db    127 GPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPG 186

Qy    311 PTTPKEPAPTTP---KEPTPTTPKEPA--PTTKEPAPTTPKEPA--------------- 349
          . |  | | |  |   ::| |     ||  | :::| |     ||
Db    187 P--PGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAHGPAGPKGAHGPAGPKGAHG 244

Qy    350 ---------PTAPK-KPAPTTP---KEPAPTTPKEPA-PTTTKEPSPTTPKEPAPTTTKS 395
                  |  ||  | | |   ::| |     || |  :::| |  |  | |  :
Db    245 PAGPKGAHGPAGPKGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP--PGAPGP----A 298

Qy    396 APTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTP---KKPAPTTPKEPAP 452
            | :::| |      |  |    |  :::| | |  | | |  |   : | |  | | |
Db    299 GPPGSRDPGP------PGAPGPAGPPGSRDPGP--PGAPGPAGPPGSRDPGP--PGAPGP 348

Qy    453 TTP---KEPAPTTTKKPA-------PTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTT 502
            |   ::| |      ||       |  | | | |  |  :    |    |  |    |
Db    349 AGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP----PGAPGPAGPPG 404

Qy    503 PEKPAPT-TPEELAPTTPEEPTPTTPEEPAPTTPKAAA-PNTPKEPAPTTP---KEPAPT 557
          | | |   |    |    :| | |  | | | | |  :  |  | | | | |   ::| |
Db    405 SRDPGPPGAPGPAGPPGSRDPGP--PGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPP 462

Qy    558 TPKEPA-------PTTPKET-APTTPKGT--------APTTLKEPAPTTPKKP-APKELA 600
            ||        |  ||    | |||          ||      |   |  | ||
Db    463 GAHGPAGPKGAHGPAGPKGAHGPAGPKGAHGPAGPKGAPGPAGPPGSRDPGPPGAPGPAG 522

Qy    601 PTTTKEPTSTTSDKPAPTTPKGTA-PTTPKEPAPTTP---KEPAPTTPKGTA-PTTLKEP 655
          |  :::|    :  | |  | |:  |  |  | |  |   ::| |    | | |   ::|
Db    523 PPGSRDPGPPGA--PGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDP 580

Qy    656 APTTPKKPAP------KELAPTTTKGPTST--TSDKPAPTTPKETAPTTPKEPAPTTPKK 707
            |  | | |       ::  |    ||     : |   |  |     |   ::| |  |
Db    581 GP--PGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP--PGA 636

Qy    708 PAPTTP----ETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALEN-----S 758
          | | |    :  ||     : |  :::|     :|  : |  | :| |
Db    637 PGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP-PGAPGPAGPPGSRDPGPPGAPGPAGPPG 695

Qy    759 PKEPGVPTTKTPAATK 774
            ::|| |    ||  |
Db    696 SRDPGPPGAHGPAGPK 711


RESULT  11
US-08-479-537A-5
; Sequence 5, Application US/08479537A
; Patent No. 5861381
;  GENERAL INFORMATION:
;    APPLICANT:  CHAMBON, Pierre
;    APPLICANT:  KIENY, Marie-Paule
;    APPLICANT:  LATHE, Richard
;    APPLICANT:  HAREUVENI, Mara
;    TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;    TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
;    NUMBER OF SEQUENCES:  5
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
;      STREET:  P.O. Box 1404
;      CITY:  Alexandria
;      STATE:  Virginia
;      COUNTRY:  United States
;      ZIP:  22313-1404
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/08/479,537A
;      FILING DATE:  07-JUN-1995
;      CLASSIFICATION:  514
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  FR 90/13101
;      FILING DATE:  23-OCT-1990
;    PRIOR APPLICATION DATA:
```

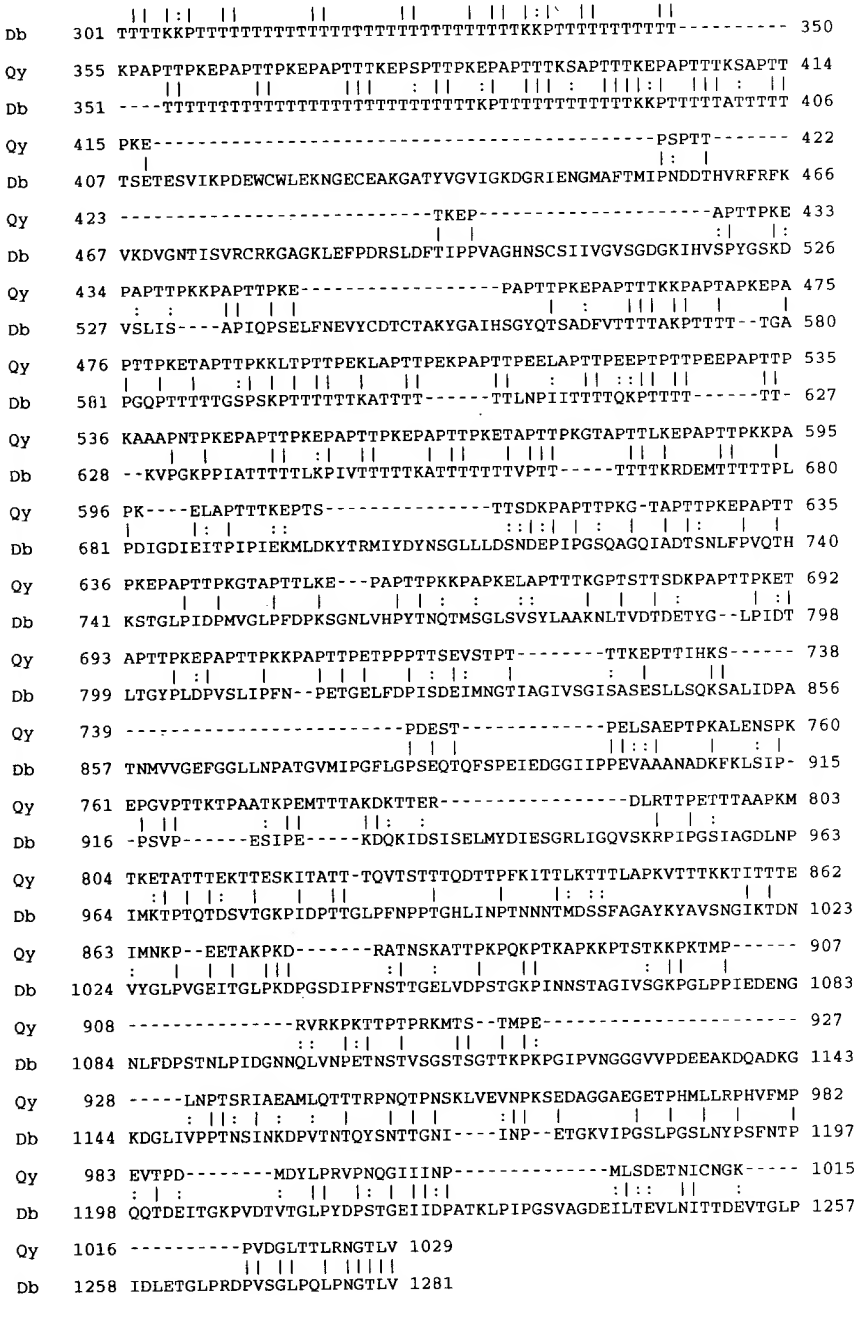
```
Qy    696 TPKEP------APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAE 749
          |  |         |: |: :  | | |  :  ::  | |    |   |:|  |
Db    650 LPIIPQKGNNIPSNLPENPSDSEVEYPRPNDNGENSNNTMKSKKNI---PNEPIPSPGDN 706

Qy    750 PTPKALENSPK-------------------EPGVPTTK 768
          |     |  ||                    || :|  :
Db    707 PYKGHEERIPKPHRSNDDYVYDNNVNKNNKDEPEIPNNE 745


RESULT   8
PCT-US92-00018-2
; Sequence 2, Application PC/TUS9200018
;  GENERAL INFORMATION:
;    APPLICANT:  Hoffman, Stephen L.
;    APPLICANT:  Charoenvit, Yupin
;    APPLICANT:  Hedstrom, Richard
;    APPLICANT:  Khusmith, Srisin
;    APPLICANT:  Rogers IV, William O.
;    TITLE OF INVENTION:  Protective malaria sporozoite surface protein
;    TITLE OF INVENTION:  immunogen and gene encoding
;    NUMBER OF SEQUENCES:  2
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  A. David Spevack
;      STREET:  NMRDC Building 1 T-12 National Naval
;      STREET:  Medical Center
;      CITY:  Bethesda
;      STATE:  MD
;      COUNTRY:  USA
;      ZIP:  20814-5044
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.24
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  PCT/US92/00018
;      FILING DATE:  19920103
;      CLASSIFICATION:  424
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Spevack, Avram D.
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (301) 295-6759
;      TELEFAX:  (301) 295-4033
;  INFORMATION FOR SEQ ID NO:  2:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  826 amino acids
;      TYPE:  AMINO ACID
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
PCT-US92-00018-2


  Query Match             7.2%;  Score 488.5;  DB 5;  Length 826;
  Best Local Similarity   26.4%;  Pred. No. 1.1e-24;
  Matches  153;  Conservative   54;  Mismatches  223;  Indels  149;  Gaps   34;

Qy    233 TPTTIKSAPTTPKEPA-PTTTKSAPTTPKEPA-PTTTKEP----APTTPKEP-APTTTKE 285
          ||  ::  |  |  |  |      |: |:||  |    :|     |  |  |  |
Db    273 TPCKVRDCPQIPIPPVIPNKIPEKPSNPEEPVNPNDPNDPNNPNNPNNPNNPNNPNNPNN 332

Qy    286 PAPTTTKSAPTTPKEP-APTTPKKP-APTTPKEP-APTTPKEP-TPTTPKEPAPTTKEPA 341
          |      :  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |:|:     |
Db    333 PNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNDPSNPNNHPK 392

Qy    342 PTTPKEPAPTAPK--KPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTT 399
            ||   |  ||  || |  |    |: | :| |       |: | :| |    : |:
Db    393 RRNPKRRNPNKPKPNKPNPNKPNPNEPSNPNKPNPN-----EPSNPNKPNP----NEPSN 443

Qy    400 TKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA-PTTPKKP-APTTPKEPAPTTPKE 457
            :| |    : |: | :|:|       |  | ||: |  |  | ||: | |  |: | |
Db    444 PNKPNP----NEPSNPNKPNPN-----EPLNPNEPSNPNEPSNPNAPSNPNE--PSNPNE 492

Qy    458 PAPTTTKKPAPTAPKEP-APTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAP 516
```

```
          |:       |  |  |  |: | |  |: |||                  |: | |  |
Db    493 PS-------NPNEPSNPNEPSNPNE--PSNPKK-----------------PSNPNE--P 523

Qy    517 TTPEEPTPTTPEEPAPTTPKAAAPNTPKEP-APTTPKEPAPTTPKEPAPTTPKETAPTTP 575
          : | |  |  | ||        : || |  |  |: |:|  |: |||  |: | |
Db    524 SNPNE--PLNPNEP-------SNPNEPSNPNEPSNPEE--PSNPKE--PSNPNE------ 564

Qy    576 KGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTT 635
                    |: |::| |:|  |:  |||    |:   |  |:   |  |  |  :
Db    565 -----------PSNPEEPNPEE--PSNPKEP----SNPEEPINPEELNPKEPSNPEESN 606

Qy    636 PKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPT 695
          ||||               |  : ||:|    : |   : |     :   |
Db    607 PKEPI------------NPEESNPKEP----INPEDNENPLIIQDEPIEPRNDSNVIPI 649

Qy    696 TPKEP------APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAE 749
          |  |          |: |: :  | | |  :  ::  | |    |   |:|  |
Db    650 LPIIPQKGNNIPSNLPENPSDSEVEYPRPNDNGENSNNTMKSKKNI---PNEPIPSPGDN 706

Qy    750 PTPKALENSPK-------------------EPGVPTTK 768
          |     |  ||                    || :|  :
Db    707 PYKGHEERIPKPHRSNDDYVYDNNVNKNNKDEPEIPNNE 745


RESULT   9
5202236-13
;Patent No. 5202236
;    APPLICANT: MAUGH, KATHY J.;ANDERSON, DAVID M.;STRAUSBERG,
;SUSAN L.;MCCANDLISS, RUSS;WEI, TENA;FILPULA, DAVID
;    TITLE OF INVENTION: METHOD OF PRODUCING BIOADHESIVE
;PROTEIN
;    NUMBER OF SEQUENCES: 39
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER: US/07/528,762
;      FILING DATE: 25-MAY-1990
;      APPLICATION NUMBER: 82,456
;      FILING DATE: 07-AUG-1987
;      APPLICATION NUMBER: 933,945
;      FILING DATE: 24-NOV-1986
;      APPLICATION NUMBER: 650,128
;      FILING DATE: 13-SEP-1984
;SEQ ID NO:13:
;      LENGTH: 652
5202236-13


  Query Match             7.0%;  Score 476.5;  DB 6;  Length 652;
  Best Local Similarity   29.7%;  Pred. No. 5.1e-24;
  Matches  221;  Conservative   71;  Mismatches  284;  Indels  169;  Gaps   45;

Qy    116 PKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKE 175
          ||:|      |:|| ||                | ::| |                | | |
Db     11 PKMTYPPTYKPKPSYPP---------------TYKSKPTY----------KPKIT---- 40

Qy    176 TQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKE----PASTTPKE 231
                  || |  |||:          |    ||:  | | |      | :  ||
Db     41 -----------YPPTYK--AKPS--------YPPTYKPKKTYPPTYKPKLTYPPTYKPKP 79

Qy    232 PTPTTIKSAPT-TPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPA--P 288
          | | || ||  ||   | | |: |: |     |  |  |    ||   | | | |  |
Db     80 SYPPTYKSKPTYKPKITYPPTYKAKPSYPPTYKPKKTYPPT-YKPKLTYPPTYKPKASYP 138

Qy    289 TTTKSAPTTPKEPAPTTPKKPAPT-TPKEPAPTT----PKEPTPTTPKEPAPTTKEPAPT 343
          | |  |: |  |:   | |    ||  ||    | |     |  |     | :  | | :|  |
Db    139 PTYKPKPSYP--PSYKTKKTYPPTYKPKLTYPPTYKPKPSYPPSYKPNKTYPPTYKPKLT 196

Qy    344 TPKEPAPTAPKKPA--PTTPKEPA--PTTPKEPAPTTTKEPSPTTP-----KEPAPTTTK 394
          |    ||   ||:  ||   :|:  ||   :|:   | :  |: |     |  | | |
Db    197 YP----PTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPSLIKAKPSYPPTYK 252

Qy    395 SAPTTTKEPA-------PTTTKSAPTTPKEPSPTTTKEPA--PTTPKEPA----PTTPKK 441
          : ||   :|         |:| |: || |    ||   :|:  ||   :|     ||   |
Db    253 AKPTYKAKPTYKAKPTYPSTYKAKPTYP----PTYKAKPSYPPTYKAKPTYKAKPTYKAK 308
```

```
Qy    415 PKE-----------------------------------------PSPTT------- 422
Db    407 TSETESVIKPDEWCWLEKNGECEAKGATYVGVIGKDGRIENGMAFTMIPNDDTHVRFRFK 466

Qy    423 ---------------------------TKEP--------------------APTTPKE 433
Db    467 VKDVGNTISVRCRKGAGKLEFPDRSLDFTIPPVAGHNSCSIIVGVSGDGKIHVSPYGSKD 526

Qy    434 PAPTTPKKPAPTTPKE-----------------PAPTTPKEPAPTTTKKPAPTAPKEPA 475
Db    527 VSLIS----APIQPSELFNEVYCDTCTAKYGAIHSGYQTSADFVTTTTAKPTTTT--TGA 580

Qy    476 PTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTP 535
Db    581 PGQPTTTTTGSPSKPTTTTTTKATTTT------TTLNPIITTTTQKPTTTT------TT- 627

Qy    536 KAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPA 595
Db    628 --KVPGKPPIATTTTTLKPIVTTTTTKATTTTTTTVPTT-----TTTTKRDEMTTTTTPL 680

Qy    596 PK----ELAPTTTKEPTS---------------TTSDKPAPTTPKG-TAPTTPKEPAPTT 635
Db    681 PDIGDIEITPIPIEKMLDKYTRMIYDYNSGLLLDSNDEPIPGSQAGQIADTSNLFPVQTH 740

Qy    636 PKEPAPTTPKGTAPTTLKE---PAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKET 692
Db    741 KSTGLPIDPMVGLPFDPKSGNLVHPYTNQTMSGLSVSYLAAKNLTVDTDETYG--LPIDT 798

Qy    693 APTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPT--------TTKEPTTIHKS------ 738
Db    799 LTGYPLDPVSLIPFN--PETGELFDPISDEIMNGTIAGIVSGISASESLLSQKSALIDPA 856

Qy    739 ------------------------PDEST-------------PELSAEPTPKALENSPK 760
Db    857 TNMVVGEFGGLLNPATGVMIPGFLGPSEQTQFSPEIEDGGIIPPEVAAANADKFKLSIP- 915

Qy    761 EPGVPTTKTPAATKPEMTTTAKDKTTER----------------DLRTTPETTTAAPKM 803
Db    916 -PSVP------ESIPE-----KDQKIDSISELMYDIESGRLIGQVSKRPIPGSIAGDLNP 963

Qy    804 TKETATTTEKTTESKITATT-TQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTE 862
Db    964 IMKTPTQTDSVTGKPIDPTTGLPFNPPTGHLINPTNNNTMDSSFAGAYKYAVSNGIKTDN 1023

Qy    863 IMNKP--EETAKPKD-------RATNSKATTPKPQKPTKAPKKPTSTKKPKTMP------ 907
Db   1024 VYGLPVGEITGLPKDPGSDIPFNSTTGELVDPSTGKPINNSTAGIVSGKPGLPPIEDENG 1083

Qy    908 --------------RVRKPKTTPTPRKMTS--TMPE---------------------- 927
Db   1084 NLFDPSTNLPIDGNNQLVNPETNSTVSGSTSGTTKPKPGIPVNGGGVVPDEEAKDQADKG 1143

Qy    928 -----LNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMP 982
Db   1144 KDGLIVPPTNSINKDPVTNTQYSNTTGNI----INP--ETGKVIPGSLPGSLNYPSFNTP 1197

Qy    983 EVTPD----------MDYLPRVPNQGIIINP-------------MLSDETNICNGK----- 1015
Db   1198 QQTDEITGKPVDTVTGLPYDPSTGEIIDPATKLPIPGSVAGDEILTEVLNITTDEVTGLP 1257

Qy   1016 ----------PVDGLTTLRNGTLV 1029
Db   1258 IDLETGLPRDPVSGLPQLPNGTLV 1281
```

```
RESULT   7
US-07-638-431-2
; Sequence 2, Application US/07638431
; Patent No. 5198535
;  GENERAL INFORMATION:
;    APPLICANT:  Hoffman, Stephen L.
;    APPLICANT:  Charoenvit, Yupin
;    APPLICANT:  Hedstrom, Richard
;    APPLICANT:  Khusmith, Srisin
;    APPLICANT:  Rogers IV, William O.
;    TITLE OF INVENTION:  Protective malaria sporozoite surface protein
;    TITLE OF INVENTION:  immunogen and gene
;    NUMBER OF SEQUENCES:  2
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  A. David Spevack
;      STREET:  NMRDC Building 1 T-12 National Naval
;      STREET:  Medical Center
;      CITY:  Bethesda
;      STATE:  MD
;      COUNTRY:  USA
;      ZIP:  20814-5044
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.24
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/07/638,431
;      FILING DATE:  19910110
;      CLASSIFICATION:  424
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Spevack, Avrom D.
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (301) 295-6759
;      TELEFAX:  (301) 295-4033
;  INFORMATION FOR SEQ ID NO:  2:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  826 amino acids
;      TYPE:  AMINO ACID
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
US-07-638-431-2
```

```
  Query Match             7.2%;  Score 488.5;  DB 1;  Length 826;
  Best Local Similarity   26.4%;  Pred. No. 1.1e-24;
  Matches  153;  Conservative   54;  Mismatches  223;  Indels  149;  Gaps   34;

Qy    233 TPTTIKSAPTTPKEPA-PTTTKSAPTTPKEPA-PTTTKEP----APTTPKEP-APTTTKE 285
Db    273 TPCKVRDCPQIPIPPVIPNKIPEKPSNPEEPVNPNDPNDPNNPNNPNNPNNPNNPNN 332

Qy    286 PAPTTTKSAPTTPKEP-APTTPKKP-APTTPKEP-APTTPKEP-TPTTPKEPAPTTKEPA 341
Db    333 PNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNDPSNPNNHPK 392

Qy    342 PTTPKEPAPTAPK--KPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTT 399
Db    393 RRNPKRRNPNKPKPNKPNPNKPNPNEPSNPNKPNPN-----EPSNPNKPNP----NEPSN 443

Qy    400 TKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA-PTTPKKP-APTTPKEPAPTTPKE 457
Db    444 PNKPNP----NEPSNPNKPNPN-----EPLNPNEPSNPNEPSNPNAPSNPNE--PSNPNE 492

Qy    458 PAPTTTKKPAPTAPKEP-APTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAP 516
Db    493 PS-------NPNEPSNPNEPSNPNE--PSNPKK-----------------PSNPNE--P 523

Qy    517 TTPEEPTPTTPEEPAPTTPKAAAPNTPKEP-APTTPKEPAPTTPKEPAPTTPKETAPTTP 575
Db    524 SNPNE--PLNPNEP-------SNPNEPSNPNEPSNPEE--PSNPKE--PSNPNE------ 564

Qy    576 KGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTT 635
Db    565 -----------PSNPEEPNPEE--PSNPKEP----SNPEEPINPEELNPKEPSNPEESN 606

Qy    636 PKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPT 695
Db    607 PKEPI------------NPEESNPKEP----INPEDNENPLIIQDEPIEPRNDSNVIPI 649
```



RESULT 6

US-08-928-361B-6
; Sequence 6, Application US/08928361B
; Patent No. 6071518
; GENERAL INFORMATION:
; APPLICANT: Petersen, Carolyn
; TITLE OF INVENTION: PEPTIDES, POLYPEPTIDES, GLYCOPROTEINS,
; TITLE OF INVENTION: THEIR FUNCTIONAL MUTANTS, VARIANTS, ANALOGS AND FRAGMENTS
; TITLE OF INVENTION: FOR TREATMENT AND DETECTION/DIAGNOSIS OF CRYPTOSPORIDIUM
; TITLE OF INVENTION: SPECIES INFECTIONS
; NUMBER OF SEQUENCES: 30
; CORRESPONDENCE ADDRESS:
; ADDRESSEE: PETERS, VERNY, JONES & BIKSA
; STREET: 385 Sherman Avenue, Suite 6
; CITY: Palo Alto
; STATE: CA
; COUNTRY: USA
; ZIP: 94306-1840
; COMPUTER READABLE FORM:
; MEDIUM TYPE: Floppy disk
; COMPUTER: IBM PC compatible
; OPERATING SYSTEM: PC-DOS/MS-DOS
; SOFTWARE: PatentIn Release #1.0, Version #1.30
; CURRENT APPLICATION DATA:
; APPLICATION NUMBER: US/08/928,361B
; FILING DATE: 12-SEP-1997
; CLASSIFICATION:
; PRIOR APPLICATION DATA:
; APPLICATION NUMBER: US 60/026,062
; FILING DATE: 13-SEP-1996
; ATTORNEY/AGENT INFORMATION:
; NAME: Verny, Hana
; REGISTRATION NUMBER: 30,518
; REFERENCE/DOCKET NUMBER: 480.76-1(HV)
; TELECOMMUNICATION INFORMATION:
; TELEPHONE: 650-324-1677
; TELEFAX: 650-324-1678
; INFORMATION FOR SEQ ID NO: 6:
; SEQUENCE CHARACTERISTICS:
; LENGTH: 1721 amino acids
; TYPE: amino acid
; STRANDEDNESS:
; TOPOLOGY: linear
; MOLECULE TYPE: protein
US-08-928-361B-6

Query Match 7.2%; Score 489; DB 3; Length 1721;
Best Local Similarity 22.1%; Pred. No. 2.4e-24;
Matches 284; Conservative 82; Mismatches 476; Indels 442; Gaps 41;

```
; Sequence 3, Application US/09103429A
; Patent No. 6187558
;  GENERAL INFORMATION:
;    APPLICANT:  Granados, Robert R
;    APPLICANT:  Wang, Ping
;    TITLE OF INVENTION:  A No. 6187558el Invertebrate Intestinal Mucin
;    TITLE OF INVENTION:  cDNA and Related Products and Methods
;    NUMBER OF SEQUENCES:  4
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  Brown, Pinnisi & Michaels, P.C.
;      STREET:  118 No. 6187558th Tioga
;      CITY:  Ithaca
;      STATE:  NY
;      COUNTRY:  USA
;      ZIP:  14850
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/09/103,429A
;      FILING DATE:  24-JUN-1998
;      CLASSIFICATION:  800
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Michaels, Christopher A
;      REGISTRATION NUMBER:  34,390
;      REFERENCE/DOCKET NUMBER:  BTI-39
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (607) 256-2000
;      TELEFAX:  (607) 256-3628
;  INFORMATION FOR SEQ ID NO:  3:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  786 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
;    HYPOTHETICAL:  NO
;    ANTI-SENSE:  NO
;    ORIGINAL SOURCE:
;      ORGANISM:  Trichoplusia ni
;      TISSUE TYPE:  peritrophic membrane
US-09-103-429A-3


  Query Match             7.3%;  Score 498.5;  DB 4;  Length 786;
  Best Local Similarity   29.1%;  Pred. No. 2.2e-25;
  Matches  200;  Conservative   25;  Mismatches  207;  Indels  255;  Gaps   32;

Qy    197 PTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAP 256
Db     93 PGPPAE-TTQAPATT----QAPTTTQAPTTTT----------QAPTTTTQ--ATTTQAP 135

Qy    257 TTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKE 316
Db    136 TTTQ--APTTTQ--ATTT---QAPTTTTQ-APTTTTQAPTTTTQ-APTTTQ--APTTTQ- 183

Qy    317 PAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTT-PKEPA-- 373
Db    184 -APTTTQAPTTITQAATTPAATTPAATTPAATTPAA-TTPAATTPGVPAPTSAPVWPPIC 241

Qy    374 ----------------------------------------------PTTTKEPSPTTPK- 386
Db    242 ELLPNGCPADFDIHLLIPHDKYCNLFYQCSNGYTFEQRCPEGLYFNPYVQRCDSPANVEC 301

Qy    387 ----EPAPTTT---------------------------------------------------- 393
Db    302 DGEISPAPPVTEGNEDEDIDIGDLLDNGCPANFEIDWLLPHGNRCDKYYQCVHGNLVERR 361

Qy    394 -------------------------KSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEP 426
Db    362 CGAGTHFSFELQQCDHIELVGCTLPGGESEEVDVDEDA-CTGWYCPTEPIEWEPLPNGCP 420

Qy    427 APTT-----PKEP----------------------APTTPKKPAPTT----------- 446
Db    421 ADFSIDHLLPHESDCGQYLQCVHGQTIARPCPGNLHFSPATQSCESPVTAGCQVFECDSD 480

Qy    447 ---PKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTP 503
Db    481 NQCTSTAAPTAAPTAAPTAAPTAAPTA--APSTVVPPATPPATAAPVPPTT---AIPT-- 533

Qy    504 EKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNT--PKEPAPTTPKEPAPTTPKE 561
Db    534 --PAPTAAPTAAPTTAAPESPTTVTVP-PTAAPTAAPTTAVPEIPITVT---SAPTAAPT 587

Qy    562 PAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPK 621
Db    588 AAPTAAPTAAPTTAVPEIPTTVTSPPTAAPTTAAP---APNTT-----------VTVPP 632

Qy    622 GTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTS 681
Db    633 TAAPTT-AAPAPNTTVTVPPTAAPTAAPPTVAH-------------APNTTAAPVTTTS 677

Qy    682 DKPAPTTPKETAPTTPKEPAPTTPKKP 708
Db    678 ---APATTPEDDDIDP--PLPNDPINP 699


RESULT   5
US-08-700-651-5
; Sequence 5, Application US/08700651B
; Patent No. 6015882
; GENERAL INFORMATION:
;  APPLICANT: PETERSEN, CAROLYN
;  APPLICANT: LEECH, JAMES
;  APPLICANT: NELSON, RICHARD, C.
;  APPLICANT: GUT, JIRI
;  TITLE OF INVENTION: VACCINES, ANTIBODIES, PROTEINS, GLYCOPROTEINS, DNAS AND RNAS
;  TITLE OF INVENTION: FOR PROPHYLAXIS AND TREATMENT OF Cryptosporidium parvum
;  TITLE OF INVENTION: INFECTIONS
;  FILE REFERENCE: 480.19-4(HV)
;  CURRENT APPLICATION NUMBER: US/08/700,651B
;  CURRENT FILING DATE: 1997-08-14
;  EARLIER APPLICATION NUMBER: 08/415,751
;  EARLIER FILING DATE: 1995-04-03
;  NUMBER OF SEQ ID NOS: 15
;  SOFTWARE: PatentIn Ver. 2.0
; SEQ ID NO 5
;   LENGTH: 1721
;   TYPE: PRT
;   ORGANISM: Cryptosporidium parvum
US-08-700-651-5


  Query Match             7.2%;  Score 489;  DB 3;  Length 1721;
  Best Local Similarity   22.1%;  Pred. No. 2.4e-24;
  Matches  284;  Conservative   82;  Mismatches  476;  Indels  442;  Gaps   41;

Qy     70 KKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPRPS 129
Db    116 RSNETK---TTEPSANTYAGVYRSN-ETKTTEPSANT---NFLLVDPKI----------N 158

Qy    130 LPPNSDTSKE--------------TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAK 174
Db    159 APCNSENSFEQGQIFDMGSKVYIPYTKCVGVKHTTTTTTTTTTTTTTTTT---------- 208

Qy    175 ETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTP 234
Db    209 --------------------------TTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 240

Qy    235 TTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSA 294
Db    241 TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 300

Qy    295 PTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPK 354
```

```
Qy    477 ------------TTPKETA-PTTP------------------------------KKL 490
Db    842 AGQIADTSNLFPVQTHKSTGLPIDPMVGLPFDPKSGNLVHPYTNQTMSGLSVSYLAAKNL 901

Qy    491 T-----------------PTTPEKLAPTTPE---------------------------- 504
Db    902 TVDTDETYGLPIDTLTGYPLDPVSLIPFNPETGELFDPISDEIMNGTIAGIVSGISASES 961

Qy    505 ----KPAPTTPEE----------LAPTT---------PEEPTPTTP--EEPAPTTPKAAA 539
Db    962 LLSQKSAPIDPATNMVVGEFGGLLNPATGVMIPGSLGPSEQTPFSPEIEDGGIIPPEVAA 1021

Qy    540 PNTPKEPAPTTPKEPAPTTPKE-----------------------------------PA 563
Db   1022 ANADKFKLSIPPSVPESIPEKDQKIDSISELMYDIESGRLIGQVSKRPIPGSIAGDLNPI 1081

Qy    564 PTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGT 623
Db   1082 MKTPTQTDSVTGKPIDPTT---GLPFNP--PTGHLINPTNNNTMDSSFAGAYKYAVSNGI 1136

Qy    624 APTT----PKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTP-----------KKPAPK-- 666
Db   1137 KTDNVYGLPVDEITGLPKDPVSDIPFNSTTGELVDPSTGKPINNYTAGIVSGKRGLPPIE 1196

Qy    667 ----ELAPTTTKGP-----------TSTT-------SDKPAPTTPKETAPTTPKEPA--- 701
Db   1197 DENGNLFDPSTKLPIDGNNQLVNPETNSTVSGSTSGSTKPKPGIPVNGGGVVPDEEAKDQ 1256

Qy    702 -----------PTTPKKPAPTTPETPPPTTSEVSTPTTTKE-PTTIHKS---PDESTPEL 746
Db   1257 ADKGKDGLIVPPTNSINKDPVTNTQYSNTTGNIINPETGKVIPGSLPGSLNYPSFNTPQQ 1316

Qy    747 SAEPTPKALE---NSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKM 803
Db   1317 TDEITGKPVDTVTGLPYDPSTGEIIDPATKLPIPGSVAGDEILTEVLNITTDEVTGLP-I 1375

Qy    804 TKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTE- 862
Db   1376 DLETGLPRDP------VSGLPQLPNGTLVD----------------PSNKKPIPGSHS 1411

Qy    863 --IMNKPEETAKPKDRATNSKATTPKPQKP-TKAP--KKPTSTKKPKTMPRVRKPKT--- 914
Db   1412 GFINGTSGEQSHEKDPSTG------KPLDPNTGLPFDEDSGSLINPETGDKLQGSHSGTF 1465

Qy    915 TPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPK-SEDA---------- 963
Db   1466 MPVPGKPQGENGGIMTPEQILEAL-------NKLPTSNEVNISPRPSSDAVPDRPTNTWW 1518

Qy    964 ---------------GGAEGETPHMLL------RPHVFMPEVTPDMDYLPRVPNQGIII 1001
Db   1519 NKISGQTFQVDGKKTIPGSAASVIHTALGTPTQTDPTTGLP--SDPSTGLPFIPGFNVLV 1576

Qy   1002 NPMLSD-------------ETNIC-----------NGKPVDGLTTL---RNGTLV-AF 1031
Db   1577 DPQTGEQIKGSVPYVSLYVKEKNIVTEAAYGLPVDPKTGFPIDPISYLPFAKNGELIDPI 1636

Qy   1032 RGHYF 1036
Db   1637 SGKYF 1641
```

RESULT 3
5202236-25
;Patent No. 5202236
; APPLICANT: MAUGH, KATHY J.;ANDERSON, DAVID M.;STRAUSBERG,
;SUSAN L.;MCCANDLISS, RUSS;WEI, TENA;FILPULA, DAVID
; TITLE OF INVENTION: METHOD OF PRODUCING BIOADHESIVE
;PROTEIN
; NUMBER OF SEQUENCES: 39
; CURRENT APPLICATION DATA:
; APPLICATION NUMBER: US/07/528,762
; FILING DATE: 25-MAY-1990
; APPLICATION NUMBER: 82,456
; FILING DATE: 07-AUG-1987
; APPLICATION NUMBER: 933,945
; FILING DATE: 24-NOV-1986
; APPLICATION NUMBER: 650,128
; FILING DATE: 13-SEP-1984
;SEQ ID NO:25:
; LENGTH: 744
5202236-25

Query Match 7.4%; Score 506.5; DB 6; Length 744;
Best Local Similarity 29.9%; Pred. No. 6.2e-26;
Matches 239; Conservative 78; Mismatches 308; Indels 175; Gaps 50;

```
Qy    116 PKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKE 175
Db     24 PKMTYPPTYKPKPSYPP----------------TYKSKPTY----------KPKIT---- 53

Qy    176 TQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKE----PASTTPKE 231
Db     54 ----------YPPTYK--AKPS--------YPPTYKPKKTYPPTYKPKLTYPPTYKPKP 92

Qy    232 PTPTTIKSAPT-TPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPA--P 288
Db     93 SYPPTYKSKPTYKPKITYPPTYKAKPSYPPTYKPKKTYPPT-YKPKLTYPPTYKPKASYP 151

Qy    289 TTTKSAPTTPKEPAPTTPKKPAPT-TPKEPAPTT----PKEPTPTTPKEPAPTTKEP--- 340
Db    152 PTYKPKPSYP--PSYKTKKTYPPTYKPKLTYPPTYKPKPSYPPSYKPKKTYPPTYKPKLT 209

Qy    341 -APTTPKEPA--PTAPKKPA--PTTPKEPA--PTTPKEPA--PTTTKEPS--PTTPKEPA 389
Db    210 YPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPS 269

Qy    390 -PTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPA--PTTPKEPA----PTTPKKP 442
Db    270 YPSTYKAKPSYT---YPSTYKAKPTYP----PTYKAKPSYPPTYKAKPTYKAKPTYKAKP 322

Qy    443 A--PTTPKEPA--PTTPKEPA--PTTTKKPA--PTAPKEPAPTTPKETAPTTPKKLTPTT 494
Db    323 SYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKVKPTYKAKPTYPS 382

Qy    495 PEKLAPTTPE--KPAPTTPEELAPTTPEEPT--PTTPEEPA-PTTPKA--AAPNT----P 543
Db    383 TYKAKPSYPPTYKAKPSYP----PTYKAKPSYPPTYKAKPSYPPTYKAKPTYPSTYKAKP 438

Qy    544 KEPAPTTPKEPAPTT----PKEPAPTTPKETAPTT--PKGTAPTTLKEPAPTTPKKPAPK 597
Db    439 SYPPTYKPKISYPPTYKAKPSYPSTYKAKSSYPPTYKAKPSYPSTYKAKSSYPPTYKAKP 498

Qy    598 ELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPA- 656
Db    499 SYPPTYKAKPTYPSTYKAKPTYK--ARPSYP----PTYKAKPS------YPPTYKAKPSY 546

Qy    657 -PTTPKKPAPKELAPTTTKGPT--STTSDKPAPTTPKETAPTTPKEPAPTTPKKPA--PT 711
Db    547 PPTYKAKPTYK-AKPTYKAKPTYPSTYKAKPSYSPTYKAKPSYP----PTYKAKPSYPPT 601

Qy    712 ---TPETPPPTTSEVSTPTTTKE----PTTIHKSPD-----ESTPELSAEPT-PKALENS 758
Db    602 YKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTYPSTYKAK 661

Qy    759 PKEPGVPTTKTPAA------TKPEMTTTAKDKTTERDLRTTPETTTAAP--------KMT 804
Db    662 PSYP--PTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTYPSTYKAKPSYPPTHKAKPT 719

Qy    805 KETATTTEKTTESKITATTT 824
Db    720 YKAKPTYPSTYKAKPTYPST 739
```

RESULT 4
US-09-103-429A-3

Matches 226; Conservative 33; Mismatches 211; Indels 342; Gaps 40;

```
Qy    190 TSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAP-----TTP 244
Db      9 TALGLVAARPEVSDAEKNPALHEPHPDXP--PAEQXXLLPXEYDCTKFYYCEYGLKFIAP 66

Qy    245 KEPAPTT------------TKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTK 292
Db     67 RDCAPGTEFKFSAQTCVHAALAGCTLPGPPAETT---QAPATTQ--APTTTQ--APTTTT 119

Qy    293 SAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTA 352
Db    120 QAPTTTTQATTTT----------QAPTTTQAPTTT----QATTTQAPTTTT-------- 156

Qy    353 PKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAP 412
Db    157 ---QAPTTTTQ-APTTTTQ-APTTTQ----------APTTTQ-APTTTQ--APTTITQAA 198

Qy    413 TTPKEPSPTTTKEPAPTTP--KEPAPTTPKKPAPTTPKEPAPTTPK-------------- 456
Db    199 TTPAATTPAAT-TPAATTPAATTPAATTPGVPAPTS----APVWPPICELLPNGCPADFD 253

Qy    457 ------------------------------------------------EPAPTTTK 464
Db    254 IHLLIPHDKYCNLFYQCSNGYTFEQRCPEGLYFNPYVQRCDSPANVECDGEISPAPPVTE 313

Qy    465 K----------------PA---------------------------------------- 467
Db    314 GNEDEDIDIGDLLDNGCPANFEIDWLLPHGNRCDKYYQCVHGNLVERRCGAGTHFSFELQ 373

Qy    468 --------------------------------PTAP--KEPAPT-----------TPKET 482
Db    374 QCDHIELVGCTLPGGESEEVDVDEDACTGWYCPTEPIEWEPLPNGCPADFSIDHLLPHES 433

Qy    483 ----------------------APTTPKKLTPTT----------PEKLAPTTPEKPAP 508
Db    434 DCGQYLQCVHGQTIARPCPGNLHFSPATQSCESPVTAGCQVFECDSDNQCTSTAAPTAAP 493

Qy    509 TTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPK 568
Db    494 TAAPTAAPTAAPTAAPSTVVPPA-TPPATAAPVPPTTAIPT----PAPTAAPTAAPTTAA 548

Qy    569 ETAPTT---PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAP 625
Db    549 PESPTTVTVPPTAAPTA----APTT----AVPEIPITVTSAPTAAPT--AAPTAAPTAAP 598

Qy    626 TTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTST-TSDKP 684
Db    599 TTAVPEIPTT-----VTSPPTAAPTT-AAPAPNT----------TVTVPPTAAPTTAAP 641

Qy    685 APTTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTP 744
Db    642 AP----NTTVTAPPTAAPTT-AAPAPNTTVTVPPTAAPTAAP-----PTVAH------AP 685

Qy    745 ELSAEPTPKALENSPKEPGVPTTKTPAATKPE 776
Db    686 NTTAAP------------VTTTSAPATT-PE 703
```

RESULT 2
US-08-928-361B-5
; Sequence 5, Application US/08928361B
; Patent No. 6071518
; GENERAL INFORMATION:
; APPLICANT: Petersen, Carolyn
; TITLE OF INVENTION: PEPTIDES, POLYPEPTIDES, GLYCOPROTEINS,
; TITLE OF INVENTION: THEIR FUNCTIONAL MUTANTS, VARIANTS, ANALOGS AND FRAGMENTS
; TITLE OF INVENTION: FOR TREATMENT AND DETECTION/DIAGNOSIS OF CRYPTOSPORIDIUM
; TITLE OF INVENTION: SPECIES INFECTIONS
; NUMBER OF SEQUENCES: 30
; CORRESPONDENCE ADDRESS:
; ADDRESSEE: PETERS, VERNY, JONES & BIKSA
; STREET: 385 Sherman Avenue, Suite 6
; CITY: Palo Alto
; STATE: CA
; COUNTRY: USA
; ZIP: 94306-1840
; COMPUTER READABLE FORM:
; MEDIUM TYPE: Floppy disk
; COMPUTER: IBM PC compatible
; OPERATING SYSTEM: PC-DOS/MS-DOS
; SOFTWARE: PatentIn Release #1.0, Version #1.30
; CURRENT APPLICATION DATA:
; APPLICATION NUMBER: US/08/928,361B
; FILING DATE: 12-SEP-1997
; CLASSIFICATION:
; PRIOR APPLICATION DATA:
; APPLICATION NUMBER: US 60/026,062
; FILING DATE: 13-SEP-1996
; ATTORNEY/AGENT INFORMATION:
; NAME: Verny, Hana
; REGISTRATION NUMBER: 30,518
; REFERENCE/DOCKET NUMBER: 480.76-1(HV)
; TELECOMMUNICATION INFORMATION:
; TELEPHONE: 650-324-1677
; TELEFAX: 650-324-1678
; INFORMATION FOR SEQ ID NO: 5:
; SEQUENCE CHARACTERISTICS:
; LENGTH: 1837 amino acids
; TYPE: amino acid
; STRANDEDNESS:
; TOPOLOGY: linear
; MOLECULE TYPE: protein
US-08-928-361B-5

Query Match 7.5%; Score 508; DB 3; Length 1837;
Best Local Similarity 21.3%; Pred. No. 1.4e-25;
Matches 295; Conservative 100; Mismatches 494; Indels 496; Gaps 48;

```
Qy     98 KVTTPDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTT 157
Db    307 KHTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 366

Qy    158 TNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPT 217
Db    367 TTTTTTTTKKPTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTKKPTT 426

Qy    218 PTTPKEPASTTPKEPTPTTIKSAPTTPKEP--APTTTKSAPTTPKEPAPTTTKEPAPTTP 275
Db    427 TTTTTTTTTTTTTTTTTTTTTTTTTTTTKKPTTTTTTTTTTTTTTTTTTTTTTTTTTTT 486

Qy    276 KEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKE----------- 324
Db    487 TTTATTTTKKPTTTTTTTTTTTTKKPTTTTTATTTTTTSETESVIKPDEWCWLEKNGECE 546

Qy    325 -------------------------PTPTT--------------------------PK 331
Db    547 AKGATYVGVIGKDGRIENGMAFTMIPNDDTHVRFRFKVKDVGNTISVRCGKGAGKLEFPD 606

Qy    332 EPAPTTKEP--------------------APTTPKEPA-PTAPKKP------------ 356
Db    607 RSLDFTIPPVAGHNSCSIIVGVSGGGKIHVSPYGSKDVSLISAPIQPCELFNEVYCDTCT 666

Qy    357 ----------------APTTPKEPAPTT---PKEPAPTTTKEPS-PTTPKEPAPTTTKS- 395
Db    667 AKYGAIHSGYQTSADFVTTTTAKPTTTTTGAPGQPTTTTTGSPSKPTTTTTTKATTTTTI 726

Qy    396 ----APTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPA 451
Db    727 LNPIITTTTQKPTTTTTTKVPGKPPIATTTTTLKPIVTTTTTKATTTTTTTVPTT----- 781

Qy    452 PTTPKEPAPTTTKKPAP-------------------------------TAPKEPAP--- 476
Db    782 TTTTKRDEMTTTTTPLPDIGDIEITPIPIEKMLDKYTRMIYDYNSGLLLDSNDEPIPGSQ 841
```

GenCore version 4.5
Copyright (c) 1993 - 2000 Compugen Ltd.

OM protein - protein search, using sw model

Run on: April 26, 2002, 16:25:33 ; Search time 61.21 Seconds
(without alignments)
466.904 Million cell updates/sec

Title: AA3
Perfect score: 6814
Sequence: 1 MAWKTLPIYLLLLLSVFVIQ.....ARAITTRSGQTLSKVWYNCP 1270

Scoring table: BLOSUM62
Gapop 10.0 , Gapext 0.5

Searched: 212252 seqs, 22503292 residues

Total number of hits satisfying chosen parameters: 212252

Minimum DB seq length: 0
Maximum DB seq length: 2000000000

Post-processing: Minimum Match 0%
Maximum Match 100%
Listing first 45 summaries

Database : Issued_Patents_AA:*
1: /cgn2_6/ptodata/2/iaa/5A_COMB.pep:*
2: /cgn2_6/ptodata/2/iaa/5B_COMB.pep:*
3: /cgn2_6/ptodata/2/iaa/6A_COMB.pep:*
4: /cgn2_6/ptodata/2/iaa/6B_COMB.pep:*
5: /cgn2_6/ptodata/2/iaa/PCTUS_COMB.pep:*
6: /cgn2_6/ptodata/2/iaa/backfiles1.pep:*

Pred. No. is the number of results predicted by chance to have a score greater than or equal to the score of the result being printed, and is derived by analysis of the total score distribution.

SUMMARIES

Result No.	Score	% Query Match	Length	DB	ID	Description
1	522	7.7	805	4	US-09-103-429A-4	Sequence 4, Appli
2	508	7.5	1837	3	US-08-928-361B-5	Sequence 5, Appli
3	506.5	7.4	744	6	5202236-25	Patent No. 5202236
4	498.5	7.3	786	4	US-09-103-429A-3	Sequence 3, Appli
5	489	7.2	1721	3	US-08-700-651-5	Sequence 5, Appli
6	489	7.2	1721	3	US-08-928-361B-6	Sequence 6, Appli
7	488.5	7.2	826	1	US-07-638-431-2	Sequence 2, Appli
8	488.5	7.2	826	5	PCT-US92-00018-2	Sequence 2, Appli
9	476.5	7.0	652	6	5202236-13	Patent No. 5202236
10	452	6.6	960	4	US-09-219-849-5	Sequence 5, Appli
11	424.5	6.2	1867	2	US-08-479-537A-5	Sequence 5, Appli
12	424.5	6.2	1867	4	US-09-083-116-5	Sequence 5, Appli
13	424.5	6.2	2035	2	US-08-479-537A-2	Sequence 2, Appli
14	424.5	6.2	2035	4	US-09-083-116-2	Sequence 2, Appli
15	419.5	6.2	2476	2	US-08-276-967-2	Sequence 2, Appli
16	419	6.1	1185	4	US-09-041-886-23	Sequence 23, Appl
17	417	6.1	829	1	US-08-642-255-132	Sequence 132, App
18	417	6.1	829	1	US-08-397-633A-53	Sequence 53, Appl
19	417	6.1	837	1	US-08-175-155-68	Sequence 68, Appl
20	417	6.1	837	1	US-08-477-509B-103	Sequence 103, App
21	417	6.1	837	1	US-08-642-255-101	Sequence 101, App
22	417	6.1	837	2	US-08-707-237A-75	Sequence 75, Appl
23	417	6.1	837	3	US-08-482-085B-103	Sequence 103, App
24	417	6.1	897	1	US-08-397-633A-50	Sequence 50, Appl
25	413.5	6.1	907	3	US-08-783-774-2	Sequence 2, Appli
26	413.5	6.1	907	5	PCT-US95-04611A-19	Sequence 19, Appl
27	404.5	5.9	408	1	US-07-609-716-65	Sequence 65, Appl
28	404.5	5.9	408	4	US-08-475-411A-65	Sequence 65, Appl
29	404.5	5.9	408	4	US-08-478-029A-65	Sequence 65, Appl
30	398.5	5.8	682	1	US-08-642-255-126	Sequence 126, App
31	398.5	5.8	682	1	US-08-397-633A-36	Sequence 36, Appl
32	381	5.6	1848	4	US-08-296-791-6	Sequence 6, Appli
33	381	5.6	1848	5	PCT-US95-10661A-6	Sequence 6, Appli
34	376	5.5	960	4	US-09-219-849-6	Sequence 6, Appli
35	375.5	5.5	1231	3	US-08-904-263A-4	Sequence 4, Appli
36	374.5	5.5	1537	1	US-08-325-267A-2	Sequence 2, Appli
37	371	5.4	761	2	US-08-707-237A-84	Sequence 84, Appl
38	371	5.4	762	1	US-08-642-255-114	Sequence 114, App
39	371	5.4	762	1	US-08-397-633A-26	Sequence 26, Appl
40	370.5	5.4	1064	1	US-08-642-255-62	Sequence 62, Appl
41	368	5.4	762	1	US-08-642-255-120	Sequence 120, App
42	368	5.4	762	1	US-08-397-633A-31	Sequence 31, Appl
43	367.5	5.4	1187	1	US-08-320-559-28	Sequence 28, Appl
44	367.5	5.4	1187	3	US-08-545-860D-28	Sequence 28, Appl
45	367.5	5.4	1187	5	PCT-US94-04496-28	Sequence 28, Appl

ALIGNMENTS

```
RESULT 1
US-09-103-429A-4
; Sequence 4, Application US/09103429A
; Patent No. 6187558
;  GENERAL INFORMATION:
;    APPLICANT:  Granados, Robert R
;    APPLICANT:  Wang, Ping
;    TITLE OF INVENTION:  A No. 6187558el Invertebrate Intestinal Mucin
;    TITLE OF INVENTION:  cDNA and Related Products and Methods
;    NUMBER OF SEQUENCES:  4
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  Brown, Pinnisi & Michaels, P.C.
;      STREET:  118 No. 6187558th Tioga
;      CITY:  Ithaca
;      STATE:  NY
;      COUNTRY:  USA
;      ZIP:  14850
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/09/103,429A
;      FILING DATE:  24-JUN-1998
;      CLASSIFICATION:  800
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Michaels, Christopher A
;      REGISTRATION NUMBER:  34,390
;      REFERENCE/DOCKET NUMBER:  BTI-39
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (607) 256-2000
;      TELEFAX:  (607) 256-3628
;  INFORMATION FOR SEQ ID NO:  4:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  805 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
;    HYPOTHETICAL:  NO
;    ANTI-SENSE:  NO
;    ORIGINAL SOURCE:
;      ORGANISM:  Trichoplusia ni
;      TISSUE TYPE:  peritrophic membrane
;
US-09-103-429A-4

  Query Match              7.7%;  Score 522;  DB 4;  Length 805;
  Best Local Similarity   27.8%;  Pred. No. 6.5e-27;
```

```
Db    412 KSPVEAKSPAEVKSPVTVKSPAEAKSPVEVKSPA-SVKSPSEAKSPAGAKSPAEAKSPVV 470

Qy    353 PKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPA----PTTTKS-----APTTTKEP 403

Db    471 AKSPAEA--KSPAGAKPPAEAKSPAEAKSPAEAKSPAEAKSPAEAKSPVEVKSPEKAKSP 528

Qy    404 APTTTKS-----APTTPKEPSPTTTKEPA----PTTPKEPAPTTPKKPAPTTPKEPAPTT 454

Db    529 VKEGAKSLAEAKSPEKAKSPVKEEIKPPAEVKSPEKAKSPMRKEAKSPEKAKTLDVKSPE 588

Qy    455 PKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEEL 514

Db    589 AKPPAKEEAKRPADIRSPEQVKSPAKEEAKSPEKEETRT--EKVAPKKEEVKSPV--EEV 644

Qy    515 APTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTT 574

Db    645 ---KAKEPPKKVEEEKTPATPKTEVKESKKDEAPKEAQKP-KAEEKEPLTEKPKDSPGEA 700

Qy    575 PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPT 634

Db    701 KKEEA----KEKKAAAPEEETPAKLGVKEEAKPKEKAEDAKA-----------KEPSKP 744

Qy    635 TPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAP 694

Db    745 SEKE----KPK-------KEEVPAAPEKKDTKE---------EKTTESKKREEKPKMEAK 784

Qy    695 TTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDE 741

Db    785 AKEED-------KGLPQEPSKPKTEKAEKSSSTDQKDSQPSEKAPED 824
```

Search completed: April 26, 2002, 16:32:37
Job time: 587 sec

```
Qy     508 PTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEP--APT 565
           ||:|   :||:|   :||:|   :||:| : :|::|:     ||: |   || .|  :|:
Db    1778 PTSP-NYSPTSPSY-SPTSPSY-SPTSP-SYSPSSPR----YTPQSPT-YTPSSPSYSPS 1828

Qy     566 TPKETAPTTPKGTAPTTLKEPAPTTPK-KPAPKELAPTTTKEPTSTTSDKPAPTTPKGTA 624
           :|   :||:|| |          ||:|   |: |  | : |   | || | :||:||  :
Db    1829 SP-SYSPTSPKYT---------PTSPSYSPSSPEYTPASPK--YSPTSPKYSPTSPK-YS 1875

Qy     625 PTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPT-STTSDK 683
           ||:|   :|||||  :||:|  |   |     ||:||       :||:   || | || |
Db    1876 PTSPTY-SPTTPKY-SPTSP--TYSPTSPVYTPTSPK------YSPTS---PTYSPTSPK 1922

Qy     684 PAPTTPKETAPTTPKEP--APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDE 741
           :||:|   :||:||    :||:|   :||:|      |  :::|           |||:
Db    1923 YSPTSP-TYSPTSPKGSTYSPTSPGY-SPTSP------TYSLTSPAI---------SPDD 1965

Qy     742 STPE 745
           |  |
Db    1966 SDEE 1969
```

```
RESULT  15
NFH_RAT
ID   NFH_RAT        STANDARD;      PRT;   831 AA.
AC   P16884; Q63368;
DT   01-AUG-1990 (Rel. 15, Created)
DT   01-FEB-1994 (Rel. 28, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   NEUROFILAMENT TRIPLET H PROTEIN (200 KDA NEUROFILAMENT PROTEIN)
DE   (NEUROFILAMENT HEAVY POLYPEPTIDE) (NF-H) (FRAGMENT).
GN   NEFH OR NFH.
OS   Rattus norvegicus (Rat).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Rattus.
OX   NCBI_TaxID=10116;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Brain;
RX   MEDLINE=89065087; PubMed=3143606;
RA   Breen K.C., Robinson P.A., Wion D., Anderton B.H.;
RT   "Partial sequence of the rat heavy neurofilament polypeptide (NF-H).
RT   Identification of putative phosphorylation sites.";
RL   FEBS Lett. 241:213-218(1988).
RN   [2]
RP   SEQUENCE OF 37-831 FROM N.A.
RX   MEDLINE=88309090; PubMed=2457365;
RA   Dautigny A., Pham-Dinh D., Roussel C., Felix J.M., Nussbaum J.L.,
RA   Jolles P.;
RT   "The large neurofilament subunit (NF-H) of the rat: cDNA cloning and
RT   in situ detection.";
RL   Biochem. Biophys. Res. Commun. 154:1099-1106(1988).
RN   [3]
RP   SEQUENCE OF 1-89 AND 243-313 FROM N.A.
RX   MEDLINE=87080760; PubMed=2878828;
RA   Robinson P.A., Wion D., Anderton B.H.;
RT   "Isolation of a cDNA for the rat heavy neurofilament polypeptide
RT   (NF-H).";
RL   FEBS Lett. 209:203-205(1986).
RN   [4]
RP   SEQUENCE OF 318-831 FROM N.A.
RX   MEDLINE=89184647; PubMed=2928342;
RA   Lieberburg I., Spinner N., Snyder S., Anderson J., Goldgaber D.,
RA   Smulowitz M., Carroll Z., Emanuel B.S., Breitner J., Rubin L.;
RT   "Cloning of a cDNA encoding the rat high molecular weight
RT   neurofilament peptide (NF-H): developmental and tissue expression in
RT   the rat, and mapping of its human homologue to chromosomes 1 and
RT   22.";
RL   Proc. Natl. Acad. Sci. U.S.A. 86:2463-2467(1989).
CC   -!- FUNCTION: NEUROFILAMENTS USUALLY CONTAIN THREE IF PROTEINS: L, M,
CC       AND H WHICH ARE INVOLVED IN THE MAINTENANCE OF NEURONAL CALIBER.
CC       NF-H HAS AN IMPORTANT FUNCTION IN MATURE AXONS THAT IS NOT
CC       SUBSERVED BY THE TWO SMALLER NF PROTEINS.
CC   -!- PTM: THERE ARE A NUMBER OF REPEATS OF THE TRIPEPTIDE K-S-P, NFH IS
CC       PHOSPHORYLATED ON A NUMBER OF THE SERINES IN THIS MOTIF. IT IS
CC       THOUGHT THAT PHOSPHORYLATION OF NFH RESULTS IN THE FORMATION OF
CC       INTERFILAMENT CROSS BRIDGES THAT ARE IMPORTANT IN THE MAINTENANCE
CC       OF AXONAL CALIBER.
CC   -!- PTM: PHOSPHORYLATION SEEMS TO PLAY A MAJOR ROLE IN THE FUNCTIONING
CC       OF THE LARGER NEUROFILAMENT POLYPEPTIDES (NF-M AND NF-H), THE
CC       LEVELS OF PHOSPHORYLATION BEING ALTERED DEVELOPMENTALLY AND
CC       COINCIDENT WITH A CHANGE IN THE NEUROFILAMENT FUNCTION.
CC   -!- SIMILARITY: BELONGS TO THE INTERMEDIATE FILAMENT FAMILY.
CC   -!- CAUTION: REF.1 SEQUENCE DIFFERS FROM THAT SHOWN FROM POSITION 783
CC       ONWARD AND IS LONGER DUE TO A FRAMESHIFT.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M37227; AAA41693.1; ALT_FRAME.
DR   EMBL; X13804; CAA32038.1; ALT_FRAME.
DR   EMBL; M21964; AAA41695.1; -.
DR   EMBL; J04517; AAA41692.1; -.
DR   PIR; A30796; A30796.
DR   PIR; A25649; A25649.
DR   PIR; B25649; B25649.
DR   PIR; S02003; S02003.
DR   InterPro; IPR001664; IF.
DR   Pfam; PF00038; filament; 1.
DR   PROSITE; PS00226; IF; 1.
KW   Intermediate filament; Coiled coil; Neurone; Phosphorylation; Repeat.
FT   NON_TER       1      1
FT   DOMAIN      276    641       51 X 3 AA TANDEM REPEATS OF K-S-P.
FT   CONFLICT    164    164       L -> I (IN REF. 2).
FT   CONFLICT    185    185       I -> S (IN REF. 2).
FT   CONFLICT    193    193       L -> T (IN REF. 2).
FT   CONFLICT    199    199       M -> T (IN REF. 2).
FT   CONFLICT    346    346       K -> N (IN REF. 1).
FT   CONFLICT    373    373       A -> V (IN REF. 1 AND 4).
FT   CONFLICT    482    482       G -> E (IN REF. 2 AND 4).
FT   CONFLICT    485    485       P -> S (IN REF. 2).
FT   CONFLICT    570    571       RK -> KE (IN REF. 2 AND 4).
FT   CONFLICT    591    591       P -> T (IN REF. 2 AND 4).
FT   CONFLICT    727    727       A -> V (IN REF. 4).
FT   CONFLICT    757    759       AAP -> GST (IN REF. 4).
FT   CONFLICT    769    769       T -> L (IN REF. 2).
FT   CONFLICT    775    775       R -> P (IN REF. 2 AND 4).
SQ   SEQUENCE   831 AA;  89486 MW;  1B0973C3F13EF768 CRC64;
```

```
  Query Match             7.3%;  Score 497;  DB 1;  Length 831;
  Best Local Similarity   27.6%;  Pred. No. 6.1e-14;
  Matches  195;  Conservative   59;  Mismatches  307;  Indels  146;  Gaps   35;

Qy     102 PDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQ 161
           |  ||  | || :  ||  :                 : ||| :   :   ::  |  |  :
Db     197 PSMST--HIKVKSEEKIKVVE-----------KSEKETVIVEEQTEEIQVTEEVTEEED 242

Qy     162 TSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEP----T 217
           |:|:  : :  :    ||     | ::  | |    |  || |     | |    :
Db     243 KEAQGEEEEEAEEGGEEAATTS-----PPAEEAASP----EKETKSPVKEEAKSPAEAKS 293

Qy     218 PTTPKEPA-STTPKEPTPTTIKSAPTTPKEPA----PTTTKSAPTTPKEPA----PTTTK 268
           |   | || : :| |     :  :|   | ||     |   || |   | ||     | | |
Db     294 PAEAKSPAEAKSPAEVKSPAVAKSPAEVKSPAEVKSPAEAKS-PAEAKSPAEVKSPATVK 352

Qy     269 EP----APTTPKEPA----PTTTKEPA----PTTTKSAPTTPKEPA----PTTPKKPA-- 310
           |    :|   | ||    |   | ||    | : || |   | ||    |   | ||
Db     353 SPGEAKSPAEAKSPAEVKSPAEAKSPAEAKSPASVKS-PGEAKSPAEAKSPAEVKSPATV 411

Qy     311 --PTTPKEPA----PTTPKEP----TPTTPKEPAPTTKEP----APTTPKEPA----PTA 352
```

```
SQ   SEQUENCE   1970 AA;  217205 MW;  6876FC25692A657E CRC64;

  Query Match             7.4%;  Score 503.5;  DB 1;  Length 1970;
  Best Local Similarity   33.3%;  Pred. No. 6.8e-14;
  Matches  181;  Conservative  101;  Mismatches  167;  Indels   95;  Gaps   47;

Qy    216 PTPTTPKEPASTTPKE-PTPTTIKSAPTTPKEPAPTTTKSAPTTPKE-----PAPTTTKE 269
          |:|     || |   :  ||     :|:      |     :|:   :     | :
Db   1507 PSPMGGISPAMTPWNQGATPAYGAWSPSVGSGMTPGAAGFSPSAASDASGFSPGYSPAWS 1566

Qy    270 PAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEP--APTTPKEPTP 327
          | | :|  | |::   |:|    |  :|  ||   |: |   ||: |  :||:|   :|
Db   1567 PTPGSPGSPGPSSPYIPSPGGAMSPSYSPTSPA-YEPRSPGGYTPQSPSYSPTSPSY-SP 1624

Qy    328 TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKE 387
          |:|   :||:   :||:|   :||:|   :||:|   :||:|           |||:|
Db   1625 TSPSY-SPTSPNYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY---------SPTSPSY 1671

Qy    388 PAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTP 447
           :||:   :||:   :||:   :||:|   |||:    :||:|   :||:|   :|||:|
Db   1672 -SPTSPSYSPTSPSY-SPTSPSYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY-SPTSP 1725

Qy    448 KEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPA 507
           :||:|   :||:    :||:|   :||:|   | ||:|   :||:|   :||:|
Db   1726 SY-SPTSPSY-SPTSPSY-SPTSPNY-SPTSPNYT-PTSP-SYSPTSP-SYSPTSPNY-T 1777

Qy    508 PTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEP--APT 565
          ||:|   :||:|   :||:|   :||:| : :|::|:      ||: |   ||  |  :|:
Db   1778 PTSP-NYSPTSPSY-SPTSPSY-SPTSP-SYSPSSPR----YTPQSPT-YTPSSPSYSPS 1828

Qy    566 TPKETAPTTPKGTAPTTLKEPAPTTPK-KPAPKELAPTTTKEPTSTTSDKPAPTTPKGTA 624
          :|   :||:|| |          ||:|   |:  |  ||: |    | || | :||:|| :
Db   1829 SP-SYSPTSPKYT---------PTSPSYSPSSPEYTPTSPK--YSPTSPKYSPTSPK-YS 1875

Qy    625 PTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPT-STTSDK 683
          ||:|   :|||||  :||:|  |   |     ||:||       :||:   || | || |
Db   1876 PTSPTY-SPTTPKY-SPTSP--TYSPTSPVYTPTSPK------YSPTS---PTYSPTSPK 1922

Qy    684 PAPTTPKETAPTTPKEP--APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDE 741
           :||:|   :||:||    :||:|   :||:|      | :::|           |||:
Db   1923 YSPTSP-TYSPTSPKGSTYSPTSPGY-SPTSP------TYSLTSPAI---------SPDD 1965

Qy    742 STPE 745
          |  |
Db   1966 SDEE 1969


RESULT  14
RPB1_MOUSE
ID   RPB1_MOUSE     STANDARD;      PRT;  1970 AA.
AC   P08775;
DT   01-NOV-1988 (Rel. 09, Created)
DT   01-MAR-1992 (Rel. 21, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   DNA-DIRECTED RNA POLYMERASE II LARGEST SUBUNIT (EC 2.7.7.6) (RPB1).
GN   POLR2A OR RPO2-1 OR RPII215.
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=87280135; PubMed=3038894;
RA   Ahearn J.M. Jr., Bartolomei M.S., West M.L., Cisek L.J., Corden J.L.;
RT   "Cloning and sequence analysis of the mouse genomic locus encoding
RT   the largest subunit of RNA polymerase II.";
RL   J. Biol. Chem. 262:10695-10705(1987).
RN   [2]
RP   SEQUENCE OF 1587-1970 FROM N.A.
RX   MEDLINE=86068017; PubMed=2999785;
RA   Corden J.L., Cadena D.L., Ahearn J.M. Jr., Dahmus M.E.;
RT   "A unique structure at the carboxyl terminus of the largest subunit
RT   of eukaryotic RNA polymerase II.";
RL   Proc. Natl. Acad. Sci. U.S.A. 82:7934-7938(1985).
RN   [3]
RP   REVISIONS, AND PRESENCE OF AN ADDITIONAL EXON.
RX   MEDLINE=92178992; PubMed=1542581;
RA   Wintzerith M., Acker J., Vicaire S., Vigneron M., Kedinger C.;
RT   "Complete sequence of the human RNA polymerase II largest subunit.";
RL   Nucleic Acids Res. 20:910-910(1992).
CC   -!- FUNCTION: DNA-DEPENDENT RNA POLYMERASE CATALYZES THE TRANSCRIPTION
CC       OF DNA INTO RNA USING THE FOUR RIBONUCLEOSIDE TRIPHOSPHATES AS
CC       SUBSTRATES.
CC   -!- CATALYTIC ACTIVITY: N NUCLEOSIDE TRIPHOSPHATE = N PYROPHOSPHATE +
CC       RNA(N).
CC   -!- SUBUNIT: RNA POLYMERASE II CONSISTS OF 12 DIFFERENT SUBUNITS.
CC   -!- SUBCELLULAR LOCATION: NUCLEAR.
CC   -!- PTM: THE TANDEM 7 RESIDUES REPEATS CAN BE HIGHLY PHOSPHORYLATED.
CC       THE PHOSPHORYLATION ACTIVATES POL2.
CC   -!- MISCELLANEOUS: THREE DISTINCT ZINC-CONTAINING RNA POLYMERASES ARE
CC       FOUND IN EUKARYOTIC NUCLEI: POLYMERASE I FOR THE RIBOSOMAL RNA
CC       PRECURSOR, POLYMERASE II FOR THE MRNA PRECURSOR, AND POLYMERASE
CC       III FOR 5S AND TRNA GENES.
CC   -!- SIMILARITY: BELONGS TO THE RNA POLYMERASE BETA' CHAIN FAMILY.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M12130; AAA40071.1; -.
DR   EMBL; M14101; AAA40071.1; JOINED.
DR   PIR; A28490; A28490.
DR   MGD; MGI:98086; Rpo2-1.
DR   InterPro; IPR000684; RNA_polII_repeat.
DR   InterPro; IPR000722; RNA_pol_A.
DR   InterPro; IPR002879; RNA_pol_A2.
DR   Pfam; PF00623; RNA_pol_A; 1.
DR   Pfam; PF01854; RNA_pol_A2; 1.
DR   PROSITE; PS00115; RNA_POL_II_REPEAT; 42.
KW   Transferase; DNA-directed RNA polymerase; Transcription; Zinc; Repeat;
KW   DNA-binding; Nuclear protein; Phosphorylation; Zinc-finger.
FT   ZN_FING      71     87       C2H2-TYPE (POTENTIAL).
FT   DOMAIN     1590   1958       CARBOXYL-TERMINAL 7-RESIDUE REPEATS.
FT   CONFLICT   1498   1498       P -> R (IN REF. 1 AND 2).
FT   CONFLICT   1499   1536       MISSING (IN REF. 1 AND 2).
SQ   SEQUENCE   1970 AA;  217175 MW;  7D76F38FD92A657E CRC64;

  Query Match             7.3%;  Score 498.5;  DB 1;  Length 1970;
  Best Local Similarity   33.1%;  Pred. No. 1.1e-13;
  Matches  180;  Conservative  101;  Mismatches  168;  Indels   95;  Gaps   47;

Qy    216 PTPTTPKEPASTTPKE-PTPTTIKSAPTTPKEPAPTTTKSAPTTPKE-----PAPTTTKE 269
          |:|     || |   :  ||     :|:      |     :|:   :     | :
Db   1507 PSPMGGISPAMTPWNQGATPAYGAWSPSVGSGMTPGAAGFSPSAASDASGFSPGYSPAWS 1566

Qy    270 PAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEP--APTTPKEPTP 327
          | | :|  | |::   |:|    |  :|  ||   |: |   ||: |  :||:|   :|
Db   1567 PTPGSPGSPGPSSPYIPSPGGAMSPSYSPTSPA-YEPRSPGGYTPQSPSYSPTSPSY-SP 1624

Qy    328 TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKE 387
          |:|   :||:   :||:|   :||:|   :||:|   :||:|           |||:|
Db   1625 TSPSY-SPTSPNYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY---------SPTSPSY 1671

Qy    388 PAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTP 447
           :||:   :||:   :||:   :||:|   |||:    :||:|   :||:|   :|||:|
Db   1672 -SPTSPSYSPTSPSY-SPTSPSYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY-SPTSP 1725

Qy    448 KEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPA 507
           :||:|   :||:    :||:|   :||:|   | ||:|   :||:|   :|||:|
Db   1726 SY-SPTSPSY-SPTSPSY-SPTSPNY-SPTSPNYT-PTSP-SYSPTSP-SYSPTSPNY-T 1777
```

```
FT   DISULFID    374    391       BY SIMILARITY.
FT   DISULFID    526    552       BY SIMILARITY.
FT   DISULFID    536    551       BY SIMILARITY.
FT   DISULFID    546    563       BY SIMILARITY.
FT   DISULFID    573    599       BY SIMILARITY.
FT   DISULFID    583    598       BY SIMILARITY.
FT   DISULFID    593    610       BY SIMILARITY.
FT   DISULFID    621    647       BY SIMILARITY.
FT   DISULFID    631    646       BY SIMILARITY.
FT   DISULFID    641    658       BY SIMILARITY.
FT   VARIANT     276    276       K -> E.
FT   VARIANT     354    354       C -> R.
FT   VARIANT     415    415       T -> A.
SQ   SEQUENCE   662 AA;  67774 MW;  F085277F1ED2FD40 CRC64;


  Query Match             7.5%;  Score 509.5;  DB 1;  Length 662;
  Best Local Similarity   28.1%;  Pred. No. 1.6e-14;
  Matches  223;  Conservative   48;  Mismatches  276;  Indels  247;  Gaps   30;

Qy    157 TTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEP 216
          ||    :  ||: | :|: : : |||:|       :| | ||  |     |   ||
Db      3 TTAAAVAATGKDTTAAAEGSAAAEKTAA-----AGEVSAPPT--AAVAATGEDATT---- 51

Qy    217 TPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPK 276
              |: |    |      ||||   || |    ||||    ||||    ||||
Db     52 --------AAATAAAETTAAAGEAPTTTTAPATTAAGKAPTTAAATAPTTAAAGAPTTAT 103

Qy    277 EPAPTTTKEPAPTTTKS-APTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPA- 334
           || |   | |||  | ||||      :|    ||||    | :  :  : :: :|
Db    104 GKAPATAAAPVPTTAASKAPTTAAAATHSTAAAAAPTTAASAAKSKERSTSSSSEEEHCH 163

Qy    335 --PTTKEPAPT---TPKEPAPTAPKKPAPTTPK----------EPAPTTPKEPAPTTTKE 379
            |: :|   :   | |:      ||     ||          :|   : :|   ||||
Db    164 VKPSKREMCGSKGITKKQ----CKKKNCCFDPKGHGGIHCFHRKPKGHSHEEHTTTTTK- 218

Qy    380 PSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTP 439
              |||| : | |||    ||||            ||| :  |||
Db    219 ---------APTTIQIATTTT---TPTTT-----------TTTTKATPTT--------- 245

Qy    440 KKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLA 499
                  ||  :  ||||            || | |  ||    ||||
Db    246 ------------TTTTKATPTTT-----------TTTKATTTTT----TPTT----- 270

Qy    500 PTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTP 559
              ||    | |||   | |||
Db    271 ---------TTTTTKATTTPTTTTTTTP------------------------------ 289

Qy    560 KEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTT 619
              |  ||   |  ||          | |          :|    :    :
Db    290 ----------TTTTTKATTTTTTTSGECKMEPSK----------REDCGYSGITESQCR 328

Qy    620 PKG-----TAPTT-----PKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELA 669
           ||     : | |              ||:     |    |    |    :|    : :
Db    329 TKGCCFDSSIPQTKWCFYTLSQVADCKVEPSQRVDCGFRGIT----ADQCRQKNCCFDSS 384

Qy    670 PTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTT 729
            : ||    :||   |  |   | |||   |  ||  |   |||      |:  :|||||
Db    385 ISGTKWCFYSTSQVAA--TKTTTTPTTTTTPTTTTTTKATTTTP------TTTTTTPTTT 436

Qy    730 KEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERD 789
            ||                | ||           || ||  | |  |||   | |
Db    437 TTTTT---------------TTKA----------TTTTPTTTTPTTTTT---KAT--- 463

Qy    790 LRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPPFKITTLKTTTLAP 849
            ||| |||  |     | |||: |       |||  |:||| ||  | ||   ||
Db    464 -TTTPTTTTTTP-----TTTTTKAT-------TTTPTTTTTTPTTTTTKATTTTPTTTTT 510

Qy    850 KVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPK----T 905
            |||| | |||    | |    |  ||        : |  :|     :|  : |    :
Db    511 TTTTTKATTTTTSGECKME----PSKRADCGYPGITESQCRSKGCCFDSSIPQTKWCFYS 566

Qy    906 MPRVRKPKTTPTPR 919
```

```
          :|:|   |  |: |
Db    567 LPQVADCKVAPSSR 580


RESULT  13
RPB1_HUMAN
ID   RPB1_HUMAN     STANDARD;      PRT;  1970 AA.
AC   P24928;
DT   01-MAR-1992 (Rel. 21, Created)
DT   01-MAR-1992 (Rel. 21, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   DNA-DIRECTED RNA POLYMERASE II LARGEST SUBUNIT (EC 2.7.7.6) (RPB1).
GN   POLR2A.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=92178992; PubMed=1542581;
RA   Wintzerith M., Acker J., Vicaire S., Vigneron M., Kedinger C.;
RT   "Complete sequence of the human RNA polymerase II largest subunit.";
RL   Nucleic Acids Res. 20:910-910(1992).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=95347616; PubMed=7622068;
RA   Mita K., Tsuji H., Morimyo M., Takahashi E., Nenoi M.,
RA   Ichimura S., Yamauchi M., Hongo E., Hayashi A.;
RT   "The human gene encoding the largest subunit of RNA polymerase II.";
RL   Gene 159:285-286(1995).
CC   -!- FUNCTION: DNA-DEPENDENT RNA POLYMERASE CATALYZES THE TRANSCRIPTION
CC       OF DNA INTO RNA USING THE FOUR RIBONUCLEOSIDE TRIPHOSPHATES AS
CC       SUBSTRATES.
CC   -!- CATALYTIC ACTIVITY: N NUCLEOSIDE TRIPHOSPHATE = N PYROPHOSPHATE +
CC       RNA(N).
CC   -!- SUBUNIT: RNA POLYMERASE II CONSISTS OF 12 DIFFERENT SUBUNITS.
CC   -!- SUBCELLULAR LOCATION: NUCLEAR.
CC   -!- PTM: THE TANDEM 7 RESIDUES REPEATS CAN BE HIGHLY PHOSPHORYLATED.
CC       THE PHOSPHORYLATION ACTIVATES POL2.
CC   -!- MISCELLANEOUS: THREE DISTINCT ZINC-CONTAINING RNA POLYMERASES ARE
CC       FOUND IN EUKARYOTIC NUCLEI: POLYMERASE I FOR THE RIBOSOMAL RNA
CC       PRECURSOR, POLYMERASE II FOR THE MRNA PRECURSOR, AND POLYMERASE
CC       III FOR 5S AND TRNA GENES.
CC   -!- SIMILARITY: BELONGS TO THE RNA POLYMERASE BETA' CHAIN FAMILY.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X63564; CAA45125.1; -.
DR   EMBL; X74874; CAA52862.1; -.
DR   EMBL; X74873; CAA52862.1; JOINED.
DR   EMBL; X74872; CAA52862.1; JOINED.
DR   EMBL; X74871; CAA52862.1; JOINED.
DR   EMBL; X74870; CAA52862.1; JOINED.
DR   PIR; S21054; S21054.
DR   MIM; 180660; -.
DR   InterPro; IPR000684; RNA_polII_repeat.
DR   InterPro; IPR000722; RNA_pol_A.
DR   InterPro; IPR002879; RNA_pol_A2.
DR   Pfam; PF00623; RNA_pol_A; 1.
DR   Pfam; PF01854; RNA_pol_A2; 1.
DR   PROSITE; PS00115; RNA_POL_II_REPEAT; 43.
KW   Transferase; DNA-directed RNA polymerase; Transcription; Zinc; Repeat;
KW   DNA-binding; Nuclear protein; Phosphorylation; Zinc-finger.
FT   ZN_FING      71     87       C2H2-TYPE (POTENTIAL).
FT   DOMAIN     1590   1958       CARBOXYL-TERMINAL 7-RESIDUE REPEATS.
FT   CONFLICT   1067   1067       W -> L (IN REF. 2).
FT   CONFLICT   1449   1449       D -> Y (IN REF. 2).
```

```
Db    186 RKPSYTPYKPKATYPPTYKPKITYP----PTYKRKPSYT-----PYKPKTTYPPTYKPKI 236

Qy    309 PAPTTPKEPAPTT----PKEPTPTT--PKEPAPTTKEPAPTTPKEPAPT-APKKPAPTTP 361

Db    237 SYPSIYKPKASYVSSYKSKKTYPPTYKPKISYPPTYKPKPSYP----PTYKPKVTYPPTY 292

Qy    362 KEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPT-----TTKEPAPTTTKSAPTTPK 416

Db    293 K-PKPSYPPTYKPKITYPPT-YKPKPSYPTPYKQKPSYPPIYKSKSSYPTSYKSKKTYPP 350

Qy    417 EPSPTTTKEPA-PTTPKEPAPTTPKKP-APT-TPK-------EPAPTTPKEPAPTTTKKP 466

Db    351 TYKPKITYPPTYKPKPSYPPSYKPKKTYSPTYKPKITYPPTYKPKPSYPPSYKPKTTYPP 410

Qy    467 APTAPKEPAPTTPKETAP-----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEE 521

Db    411 T-YKPKISYPPTYKPKASYVSSYKSKKTYPPTYKPKISYPPTYKPKPSYPPTYKPKITYP 469

Qy    522 PT-PTTPEEPAPTTPKAAAPNTPK-EPAPTTPKEPAPTTP-----KEPAPTT--PKETAP 572

Db    470 PTYKPKPSYPPTYKPKITYPPTYKRKPSYPTPYKQKPSYPPIYKSKSSYPTSYKSKKTYP 529

Qy    573 TT--PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPT-TPKGTAPTTPK 629

Db    530 PTYKPKITYPPTYK------PKPSYPPSYKPKTTYPPTYKPKIRYPPTYKPKASYPPTYK 583

Qy    630 EPAPTTPKEPAPTTPKGTAPTTLKEPAPTTP----KKPAPKELAPTTTKGPTSTTSDKPA 685

Db    584 --------------PKITYPPTYK-PKPSYPTPYKQKPSYPPIYKSKSSYPTAYKSKKTY 628

Qy    686 PTT--PKETAPTTPKEPAPTTPK--KPAPTTPET-------PPPTTSEVSTPTTTKEPTT 734

Db    629 PPTYKPKITYPPTYK-PKPSYPPSYRPKITYPPTYKPKKSYPQAYKSKGSYPPSYQPKKT 687

Qy    735 IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKT----TERDL 790

Db    688 Y---PPSYKPKKTYPPTYK-----PKISYPPTYK----TKPSYPASYKRKTSYPPTYKPK 735

Qy    791 RTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPK 850

Db    736 ISYPSTYKAKP-----SYPPTYKPKPSYASSYKPKIRYPPTYKPKPSYASSYK-----PK 785

Qy    851 VTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ---KPTKAPKKP-TSTKKPK-T 905

Db    786 I--------------RYPPTYKPKP----SYASSYKPKIRYPPTYKPKPSYASSYKPKIT 827

Qy    906 MPRVRKPKTT--PTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQ 947

Db    828 YPPTYKPKISYPPTYKPKITYPPTYKPKISYPPAYKPKISYPSQ 871
```

```
RESULT  12
MUC1_XENLA
ID   MUC1_XENLA     STANDARD;      PRT;   662 AA.
AC   Q05049;
DT   01-OCT-1994 (Rel. 30, Created)
DT   01-OCT-1994 (Rel. 30, Last sequence update)
DT   01-OCT-1994 (Rel. 30, Last annotation update)
DE   INTEGUMENTARY MUCIN C.1 (FIM-C.1) (FRAGMENT).
OS   Xenopus laevis (African clawed frog).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Amphibia; Batrachia; Anura; Mesobatrachia; Pipoidea; Pipidae;
OC   Xenopodinae; Xenopus.
OX   NCBI_TaxID=8355;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Skin;
RX   MEDLINE=93077556; PubMed=1447205;
RA   Hauser F., Hoffmann W.;
RT   "P-domains as shuffled cysteine-rich modules in integumentary mucin
RT   C.1 (FIM-C.1) from Xenopus laevis. Polydispersity and genetic
RT   polymorphism.";
RL   J. Biol. Chem. 267:24620-24624(1992).
```

```
CC   -!- FUNCTION: COULD BE INVOLVED IN DEFENSE AGAINST MICROBIAL
CC       INFECTIONS. PROTECTS THE EPITHELIA FROM EXTERNAL ENVIRONMENT.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- ALTERNATIVE PRODUCTS: A NUMBER OF DIFFERENT FORMS OF THE PROTEIN
CC       MAY BE PRODUCED BY ALTERNATIVE SPLICING.
CC   -!- TISSUE SPECIFICITY: SKIN.
CC   -!- PTM: EXTENSIVELY O-GLYCOSYLATED.
CC   -!- SIMILARITY: CONTAINS 6 P-TYPE (TREFOIL) DOMAINS.
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; L02115; AAA74725.1; -.
DR   PIR; A45155; A45155.
DR   HSSP; P04002; 1WFA.
DR   InterPro; IPR000519; P_trefoil.
DR   Pfam; PF00088; trefoil; 6.
DR   SMART; SM00018; P; 6.
DR   PROSITE; PS00025; P_TREFOIL; 6.
KW   Repeat; Amphibian skin; Glycoprotein; Alternative splicing.
FT   NON_TER       1      1
FT   DOMAIN       81    144       8 X 8 AA APPROXIMATE TANDEM REPEATS,
FT                                ALA/THR-RICH.
FT   REPEAT       81     88       1-1.
FT   REPEAT       89     96       1-2.
FT   REPEAT       97    104       1-3.
FT   REPEAT      105    112       1-4.
FT   REPEAT      113    120       1-5.
FT   REPEAT      121    128       1-6.
FT   REPEAT      129    136       1-7.
FT   REPEAT      137    144       1-8.
FT   DOMAIN      161    202       P-TYPE 1.
FT   DOMAIN      218    301       8 X APPROXIMATE TANDEM REPEATS, THR-RICH.
FT   REPEAT      218    224       2-1.
FT   REPEAT      225    239       2-2.
FT   REPEAT      240    249       2-3.
FT   REPEAT      250    259       2-4.
FT   REPEAT      260    275       2-5.
FT   REPEAT      276    287       2-6.
FT   REPEAT      288    294       2-7.
FT   REPEAT      295    301       2-8.
FT   DOMAIN      306    347       P-TYPE 2.
FT   DOMAIN      353    394       P-TYPE 3.
FT   DOMAIN      402    522       12 X APPROXIMATE TANDEM REPEATS,
FT                                THR-RICH.
FT   REPEAT      402    411       3-1.
FT   REPEAT      412    419       3-2.
FT   REPEAT      420    431       3-3.
FT   REPEAT      432    443       3-4.
FT   REPEAT      444    453       3-5.
FT   REPEAT      454    460       3-6.
FT   REPEAT      461    472       3-7.
FT   REPEAT      473    479       3-8.
FT   REPEAT      480    491       3-9.
FT   REPEAT      492    498       3-10.
FT   REPEAT      499    515       3-11.
FT   REPEAT      516    522       3-12.
FT   DOMAIN      525    566       P-TYPE 4.
FT   DOMAIN      572    613       P-TYPE 5.
FT   DOMAIN      620    661       P-TYPE 6.
FT   DISULFID    162    188       BY SIMILARITY.
FT   DISULFID    172    187       BY SIMILARITY.
FT   DISULFID    182    199       BY SIMILARITY.
FT   DISULFID    307    333       BY SIMILARITY.
FT   DISULFID    317    332       BY SIMILARITY.
FT   DISULFID    327    344       BY SIMILARITY.
FT   DISULFID    354    380       BY SIMILARITY.
FT   DISULFID    364    379       BY SIMILARITY.
```

```
FT   CONFLICT    534    534       A -> E (IN REF. 2).
FT   CONFLICT    699    699       I -> T (IN REF. 2).
FT   CONFLICT    703    703       V -> L (IN REF. 2).
FT   CONFLICT    721    721       D -> E (IN REF. 2).
SQ   SEQUENCE   865 AA;  84781 MW;  2110417E0B0E7CFE CRC64;
```

```
  Query Match             7.8%;  Score 530;  DB 1;  Length 865;
  Best Local Similarity   26.2%;  Pred. No. 2.9e-15;
  Matches  233;  Conservative   73;  Mismatches  433;  Indels  152;  Gaps   41;

Qy    188 APTSKVLAKP-TPK---AETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTT 243
          :| |  :| | ||     |       ||   |       |  | : ||  | ||     | |
Db      8 SPVSAPVAAPVTPSAVAAPVQVVSPAAVAP------APAAPIAVTPVAPPPTLASVQPAT 61

Qy    244 PKEPAPTTTKSAPTTP-KEPAPTTTKEPAPTTPKEPAPTTTKEPAPT----TTKSAPTTP 298
            |||    :|  ||    ||      | |     || :    |          |||  |
Db     62 VTIPAPAPIAAASVTPVASVAPPVVAAPTP-----PAASPVSTPVAVAQIPVAVSAPVAP 116

Qy    299 KEPAPTTPKKPAP-------TTPKEPAPTTPKEPTP-------TTPKEPAPTTKEPAPTT 344
            |  ||    |       | |   :  ||   ||       : |  || ||   |
Db    117 PVAATPTPVVQIPVAAPVIATPPVAASAPTPAAVTPVISPVIASPPVVPANTTVPVAAPV 176

Qy    345 PKEPA--PTAPKKPAPTTPKEPAPTTPKEPAPTTTKE-PSPTTPKEPAP-----TTTKSA 396
          ||  |      ||      ||   : |||    | |  | |:  ||     :   :
Db    177 AAVPAAVPVVAPVLAPAVAPAVAPVVAETPAPPPVAEIPVATIPECVAPLIPEVSVVATK 236

Qy    397 PTTTKEP---APTTTKSAPTTPKEPSPTTTKEPA-PTTPKEPAPTTPKKPAPTTPKEPAP 452
          |    ||   ||  |::    |   ||  :  ||  |    |   : : |           ||
Db    237 PLAAAEPVVVAPPATETPVVAPAAASPHVSVAPAVETAVVAPVSASTEPPVAAATLTTAP 296

Qy    453 TTP--------KEPAPTTTKKPAPTAPKEP---APTTPKET--------APTTPKKLTPT 493
           ||        : |  |     ||   ||   ||    ||          | |||  : |
Db    297 ETPALAPVVAESQVAANTVVATPPTPAPEPETIAPPVVAETPEVASVAVAETTPPVVPPV 356

Qy    494 TPEKL-APTTPEKPAPTTPEELAPTTPEEPTPTTPEEP---APTTPKAAAPNTPKEPAPT 549
          | : ||     | | |   || | |:      || |   ||:   :|   || : ||
Db    357 AAESIPAPVVATTPVPAT---LAVTDPDVTASAVPELPPVIAPSPVPSAVAETPVDLAPP 413

Qy    550 TPKEPAPTTPKEPAPTTPKETAPTTP-KGTAPTTLKE----------PAPTTPKKPAPKE 598
            |   || |    |  | ||   :|| |:             ||:     ||
Db    414 V----LPPVAAEPVPAVVAEETPETPAPASAPVTIAALDIPEVAPVIAAPSDAPAEAPSA 469

Qy    599 LAPTTTKEPTSTTSDKPAPTTPKGTAPTTP------KEPAPTTPKEP--APTTPKGTAPT 650
          || :  |  ||:  | | |    || |       | |  ||  |    ||    |
Db    470 AAPIVSTPP--TTASVPETTAPPAAVPTEPIDVSVLSEAAIETPVAPPVEVTTEVAVADV 527

Qy    651 TLKEPA------PTTPKKPAPKELAPTTT----KGPTSTTSDKPAPTT----PKETAPTT 696
          | |      |  |  | |  |  ||:    :   ||:|  | | |    | |   :
Db    528 APPEAAADLIIEPVEPPAPIPDLLEQTTSVPAVEAAESTSS--PIPETSLPPPNEAVASP 585

Qy    697 PKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPD----------ESTPEL 746
          || |  :| |     | |: :  : |   : |  | :: |              | ||
Db    586 EVAVAPITAPEPIP----EPEPSLATPTEPIPVEAPVVIQEAVDAVEVPVTETSTSIPET 641

Qy    747 SAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTP--ETTTAAPKMT 804
          : |  |:|:     :| :  |: |  |:       |     :: ||| |  ||| ::: 
Db    642 TVE-FPEAVAEKVLDPAI--TEAPVTTQEPDVANINDGAPATEI-TTPAVEIVTAAAEVS 697

Qy    805 KETATTTEKTTESKITATTTQVTST------TTQDTTPFKITTLKTTTLA-PKVTTTKKT 857
              :     :|       | |        | | | :    : :  | | ::
Db    698 DIAIPVIDPPVPQEIAVAEIPETDTKPAEVIVEQSTIPIEAPVPEVSKYAEPVISEAPAA 757

Qy    858 ITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPT 917
                : |: |:        : |  |  : ||   : |  :  |      |  | ||
Db    758 EVPITAGDNPDNTSVGISEVVPTIAEKPVEEVPTS--EIPEQSSSPS--DSVPVAKITPL 813

Qy    918 PRKMTSTMPELNPTSRIAEAM---LQTTTRPNQTPNSKLVEVN-----PKS 960
          | : :|   |   :   :|:   |:     ||    :| |:      |||
Db    814 LRDLQTTDVSLLAIAATLDAIGEKLKDQKARNQQVMDRLCEIEKILGPPKS 864
```

RESULT 11

```
FP1_MYTCO
ID   FP1_MYTCO      STANDARD;      PRT;   872 AA.
AC   Q25434;
DT   30-MAY-2000 (Rel. 39, Created)
DT   30-MAY-2000 (Rel. 39, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ADHESIVE PLAQUE MATRIX PROTEIN PRECURSOR (FOOT PROTEIN 1) (MCFP1).
GN   FP1.
OS   Mytilus coruscus (Sea mussel).
OC   Eukaryota; Metazoa; Mollusca; Bivalvia; Pteriomorphia; Mytiloida;
OC   Mytiloidea; Mytilidae; Mytilus.
OX   NCBI_TaxID=42192;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Foot;
RX   MEDLINE=96394686; PubMed=8798340;
RA   Inoue K., Takeuchi Y., Takeyama S., Yamaha E., Yamazaki F., Odo S.,
RA   Harayama S.;
RT   "Adhesive protein cDNA sequence of the mussel Mytilus coruscus and
RT   its evolutionary implications.";
RL   J. Mol. Evol. 43:348-356(1996).
CC   -!- FUNCTION: PROVIDES ADHESIVENESS TO THE MUSSEL'S FOOT. MUSSELS
CC       PRODUCE ONE OF THE STRONGEST WATER INSOLUBLE GLUE. THE MUSSEL'S
CC       ADHESIVE IS A BUNDLE OF THREADS, CALLED A BYSSUS, FORMED BY A
CC       FIBROUS COLLAGENOUS CORE COATED WITH ADHESIVE PROTEINS.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- TISSUE SPECIFICITY: PRODUCED BY THE BYSSAL GLAND.
CC   -!- DOMAIN: ALMOST EXCLUSIVELY COMPOSED OF REPEATS OF A DECAPEPTIDE.
CC   -!- PTM: PROLINES IN THE REPEAT ARE HYDROXYLATED (SINGLE OR DI-) AND
CC       ALSO TYROSINE (THUS PRODUCING DOPA = 3,4-DIHYDROXYPHENYLALANINE).
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; D63777; BAA09850.1; -.
DR   InterPro; IPR002964; Adhesive_plaq.
DR   InterPro; IPR002965; P_rich_extensn.
DR   PRINTS; PR01216; ADHESIVEI.
DR   PRINTS; PR01217; PRICHEXTENSN.
KW   Signal; Repeat; Hydroxylation.
FT   SIGNAL        1     20       POTENTIAL.
FT   CHAIN        21    872       ADHESIVE PLAQUE MATRIX PROTEIN.
FT   DOMAIN       21     41       NONREPETIVE LINKER.
FT   DOMAIN      124    872       TANDEM REPEATS OF Y-K-[PS]-K-[IP]-[ST]-Y-
FT                                P-[PST]-[ST].
FT   DOMAIN      184    192       NONAPEPTIDE 1.
FT   DOMAIN      213    221       NONAPEPTIDE 2.
SQ   SEQUENCE   872 AA;  101677 MW;  98CC70D7C75FF3C4 CRC64;
```

```
  Query Match             7.6%;  Score 518.5;  DB 1;  Length 872;
  Best Local Similarity   29.2%;  Pred. No. 8.6e-15;
  Matches  276;  Conservative   80;  Mismatches  407;  Indels  181;  Gaps   54;

Qy     84 VVDEAGSGLDNGDFKVTTPDTST--TQH-------NKVSTSPKITTAKPINPRPSLPPNS 134
          |   | ||   | :| | | :    ::|       ||: | | |  :|   | |  |
Db     29 VYGSAYSGASAGAYK-TLPGSHPYGSKHVPVYKPMNKIPT-PYI--SKKSYPAPYKPKGY 84

Qy    135 DTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVL 194
            :|    |   :|           | :|    : |  | : ::    | |      |:
Db     85 YPTKRYQPTYGSKTNYPPIYKPIAKKLSSYKAIKTTYPAYKAKTSYPPSYK-----HKIT 139

Qy    195 AKPTPKAETTTKGPALTTPKEPT-PTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTK 253
            || |       | :| |  ||    |  | |  ||   | | |     ||   | | |
Db    140 YPPTYK-------PKITYP--PTYKQKPSYPPSYKPKTTYPPTYK-----PKITYPPTYK 185

Qy    254 SAPT-TPKEPAPT--TTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTT--PKK 308
            |: || :|  |   | :|  | |    ||  ::|: |     |  ||   | |  ||
```

```
CC   -!- SUBCELLULAR LOCATION: ATTACHED TO THE MEMBRANE BY A GPI-ANCHOR
CC       (POSSIBLE).
CC   -!- DEVELOPMENTAL STAGE: MAXIMAL ACTIVITY IN TRYPOMASTIGOTES, MINIMUM
CC       IN EPIMASTIGOTES AND NO DETECTABLE IN AMASTIGOTES.
CC   -!- MISCELLANEOUS: THE VARIABLE LENGTHS OF THE LONG TANDEM REPEAT
CC       DOMAIN COULD ACCOUNT IN PART FOR THE POLYMORPHISM OF THE TNCA
CC       PROTEIN.
CC   -!- SIMILARITY: BELONGS TO FAMILY 33 OF GLYCOSYL HYDROLASES.
CC   -!- SIMILARITY: CONTAINS 3 BNR REPEATS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M61732; AAA30255.1; -.
DR   PIR; JH0557; JH0557.
DR   HSSP; P29768; 1DIL.
DR   InterPro; IPR002860; BNR.
DR   Pfam; PF02012; BNR; 2.
KW   Hydrolase; Glycosidase; Glycoprotein; Repeat; GPI-anchor;
KW   Phosphorylation.
FT   DOMAIN       1    457       CYS-RICH.
FT   REPEAT      23     34       BNR 1.
FT   REPEAT     163    174       BNR 2.
FT   REPEAT     209    220       BNR 3.
FT   DOMAIN     458    588       FIBRONECTIN TYPE-III.
FT   DOMAIN     589   1120       44 X 12 AA TANDEM REPEATS, LTR DOMAIN.
FT   CARBOHYD   342    342       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   394    394       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  1125   1125       N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE  1162 AA;  120032 MW;  07049221897C6A40 CRC64;


  Query Match             7.8%;  Score 530.5;  DB 1;  Length 1162;
  Best Local Similarity   29.4%;  Pred. No. 3.6e-15;
  Matches  177;  Conservative  100;  Mismatches  212;  Indels  113;  Gaps   34;

Qy    217 TPTTPKE-PASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTP 275
          ||:|| :  | :|| |  :  || :|| || ::    |:|| : :  |    |:||
Db    599 TPSTPADSSAHSTPSTPVDS---SAHSTPSTPADSSAHGTPSTPVDSSAHGT----PSTP 651

Qy    276 KE-PAPTTTKEPAPTTTKSAPTTPKE-PAPTTPKKP-------APTTPKEPA----PTTP 322
           :  |  |   |  ::  | |:|| :  | :|| |       ||:|| : :    |:||
Db    652 ADSSAHGTPSTPVDSSAHSTPSTPVDSSAHSTPSTPVDSSAHGAPSTPADSSAHGTPSTP 711

Qy    323 KEPTP-TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPS 381
           : :   ||  || ::    |:|| : :       | :||  || ::       |     :
Db    712 VDSSAHGTPSTPADSSAHSTPSTPADSS------AHSTPSTPADSSAHSTPSTPVDSSA 764

Qy    382 PTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA----PT 437
            ||  || ::  | |:|   || ::    |:|| : |  :|    |:|| : :    |:
Db    765 HGTPSTPADSSAHSTPST---PADSSAHGTPSTPVDSSAHST----PSTPVDSSAHGTPS 817

Qy    438 TP-KKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKL----TP 492
           ||    | :||  |  ::       |     | : |  || ::    | |:||       ||
Db    818 TPVDSSAHSTPSTPVDSSAHGTPSTPVDSSAHSTPSTPADSSAHST-PSTPADSSAHGTP 876

Qy    493 TTP-EKLAPTTPEKPAPTTPEELAPTTPEE----PTPTTPEEPA----PTTP-KAAAPNT 542
           :|| :  | :||  || ::     |:|| :     ||:|| : :   |:||  ::|  |
Db    877 STPVDSSAHSTPSTPADSSAHS-TPSTPVDSSAHSTPSTPADSSAHGTPSTPVDSSAHGT 935

Qy    543 PKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPT 602
          |  || ::    | :||  || ::   | |:||  ::       | :||  |       |
Db    936 PSTPADSS----AHSTPSTPADSSAHST-PSTPADSS------AHSTPSTPVDSSAHST 983

Qy    603 TTKEPTSTTSDKPAPTTPKGTAPTTPKE-PAPTTPKEPAPTTPKGTAPTTLKEPAPTTPK 661
                || :|  | :|     |:|| :  | :||  || ::   |  | :   | :||
Db    984 -----PSTPADSSAHST-----PSTPADSSAHSTPSTPADSSAHSTPSTPVDSSAHSTPS 1033

Qy    662 KPAPKELAPTTTKGPTSTTSDKPAPTTPK----ETAPTTPKEPAPTTPKKPAPTTPETPP 717
            |     | ::  |  || :|  | :||      :| :||  || ::    |  || ||
Db   1034 TP-----ADSSAHGTPSTPADSSAHSTPSTPVDSSAHSTPSTPADSS----AHGTPSTPA 1084

Qy    718 PTTSEVSTPTTTKEPTTIHKSPDESTP-ELSAEPTPKALENSPKEPGVPTTKTPAATKPE 776
           :::  |||:|  : :: | :|  ||| : ||  ||                 ||| :
Db   1085 DSSAH-STPSTPAD-SSAHGTP--STPADSSAHSTP--------------STPAGSSAN 1125

Qy    777 MT 778
           |
Db   1126 GT 1127


RESULT  10
CPN_DROME
ID   CPN_DROME      STANDARD;      PRT;   865 AA.
AC   Q02910;
DT   01-OCT-1993 (Rel. 27, Created)
DT   01-OCT-1993 (Rel. 27, Last sequence update)
DT   01-FEB-1994 (Rel. 28, Last annotation update)
DE   CALPHOTIN.
GN   CPN OR CAP.
OS   Drosophila melanogaster (Fruit fly).
OC   Eukaryota; Metazoa; Arthropoda; Tracheata; Hexapoda; Insecta;
OC   Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha;
OC   Ephydroidea; Drosophilidae; Drosophila.
OX   NCBI_TaxID=7227;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=CANTON-S;
RX   MEDLINE=93165729; PubMed=8094559;
RA   Martin J.H., Benzer S., Rudnicka M., Miller C.A.;
RT   "Calphotin: a Drosophila photoreceptor cell calcium-binding protein.";
RL   Proc. Natl. Acad. Sci. U.S.A. 90:1531-1535(1993).
RN   [2]
RP   SEQUENCE FROM N.A.
RC   STRAIN=CANTON-S;
RX   MEDLINE=93165730; PubMed=8434015;
RA   Ballinger D.G., Xue N., Harshman K.D.;
RT   "A Drosophila photoreceptor cell-specific protein, calphotin, binds
RT   calcium and contains a leucine zipper.";
RL   Proc. Natl. Acad. Sci. U.S.A. 90:1536-1540(1993).
CC   -!- FUNCTION: MIGHT FUNCTION AS A CALCIUM-SEQUESTERING "SPONGE" TO
CC       REGULATE THE AMOUNT OF FREE CYTOPLASMIC CALCIUM. IT BINDS 0.3 MOL
CC       OF CA+2 PER MOL OF PROTEIN.
CC   -!- SUBUNIT: HOMODIMER (PROBABLE).
CC   -!- SUBCELLULAR LOCATION: CYTOPLASMIC; HYPODENSE COMPARTMENT.
CC   -!- TISSUE SPECIFICITY: SOMA AND AXONS OF PHOTORECEPTOR CELLS OF
CC       COMPOUND EYES AND OCELLI.
CC   -!- DEVELOPMENTAL STAGE: EXPRESSED EARLY IN PHOTORECEPTOR CELL
CC       DEVELOPMENT.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; L02111; AAA28405.1; -.
DR   EMBL; L05080; AAA28420.1; -.
DR   PIR; A47282; A47282.
DR   FlyBase; FBgn0010218; Cpn.
KW   Calcium-binding.
FT   CONFLICT     36     36       A -> AVAPAVVA (IN REF. 2).
FT   CONFLICT     43     43       I -> T (IN REF. 2).
FT   CONFLICT     64     64       I -> V (IN REF. 2).
FT   CONFLICT     76     76       T -> A (IN REF. 2).
FT   CONFLICT    100    100       P -> PP (IN REF. 2).
FT   CONFLICT    126    127       VQ -> AP (IN REF. 2).
FT   CONFLICT    154    154       I -> V (IN REF. 2).
FT   CONFLICT    160    160       S -> T (IN REF. 2).
```

```
CC       TO 716 AND IS SHORTER DUE TO FRAMESHIFTS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M24496; AAA39813.1; -.
DR   EMBL; M23349; AAA39813.1; JOINED.
DR   EMBL; M24494; AAA39813.1; JOINED.
DR   EMBL; M24495; AAA39813.1; JOINED.
DR   EMBL; M35131; AAA39809.1; ALT_FRAME.
DR   EMBL; Z31012; CAA83229.1; -.
DR   PIR; JT0368; QFMSH.
DR   PIR; A43778; A43778.
DR   MGD; MGI:97309; Nfh.
DR   InterPro; IPR001664; IF.
DR   Pfam; PF00038; filament; 1.
DR   PROSITE; PS00226; IF; 1.
KW   Intermediate filament; Coiled coil; Neurone; Phosphorylation;
KW   Repeat.
FT   DOMAIN        1     97       HEAD.
FT   DOMAIN       98    408       ROD.
FT   DOMAIN      409   1087       TAIL.
FT   DOMAIN      436    517       GLU-RICH (ACIDIC).
FT   DOMAIN      519    886       50 X 6 AA TANDEM REPEATS OF K-S-P-A-E-A.
FT   DOMAIN      887   1087       GLU/LYS-RICH.
FT   DOMAIN       98    129       COIL 1A.
FT   DOMAIN      130    141       LINKER 1.
FT   DOMAIN      142    239       COIL 1B.
FT   DOMAIN      240    261       LINKER 12.
FT   DOMAIN      262    283       COIL 2A.
FT   DOMAIN      284    287       LINKER 2.
FT   DOMAIN      288    408       COIL 2B.
FT   CONFLICT    133    133       K -> QA (IN REF. 2 AND 3).
FT   CONFLICT    199    199       A -> AR (IN REF. 2 AND 3).
FT   CONFLICT    281    281       S -> T (IN REF. 2 AND 3).
FT   CONFLICT    492    492       L -> G (IN REF. 2 AND 3).
FT   CONFLICT    551    551       P -> PREAKSP (IN REF. 3).
FT   CONFLICT    689    712       MISSING (IN REF. 3).
FT   CONFLICT    714    714       G -> A (IN REF. 3).
FT   CONFLICT    814    814       V -> M (IN REF. 2 AND 3).
FT   CONFLICT    843    843       T -> N (IN REF. 2 AND 3).
SQ   SEQUENCE   1087 AA;  116612 MW;  57BAC76A38ED1CB9 CRC64;
```

```
  Query Match             7.8%;  Score 533;  DB 1;  Length 1087;
  Best Local Similarity   27.5%;  Pred. No. 2.7e-15;
  Matches  216;  Conservative   81;  Mismatches  330;  Indels  158;  Gaps   38;

Qy      86 DEAGSGLDNGDFKVT-----TPDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTSKET 140
           :|   |      | :|      |  ||  | || :   |   :                 : |||
Db     404 EECRIGFGPSPFSLTEGLPKIPSIST--HIKVKSEEMIKVVE-----------KSEKET 449

Qy     141 SLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPK 200
           :   :   :   | |  :     |:|   : :   :  |:  |   :| ::  | |  :
Db     450 VIVEGQTEEIRVTEGVTEEEDKEAQGQEGEEAEEGEEKEEEELAAATSPPAEEAASPEKE 509

Qy     201 AETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPK 260
           ::  |  | :  :  :|   | ||    :  :|   || |   | |    : : | :|
Db     510 TKSRVKEEAKSPGEAKSPGEAKSPA----EAKSPGEAKS-PGEAKSPGEAKSPAEPKSPA 564

Qy     261 EP-APTTTKEPAPTTPKEPAPTTTKEP----APTTTKSAPTTPKEPA----PTTPKKPA- 310
           || :|   | ||   || ||  | | |    :|:  || |   | ||     |   | ||
Db     565 EPKSPAEAKSPA--EPKSPA--TVKSPGEAKSPSEAKS-PAEAKSPAEAKSPAEAKSPAE 619

Qy     311 ---PTTPKEPA----PTTPKEP----TPTTPKEPAPTTKEPAPTTPKEPA----PTAPKK 355
             |   | ||    | | | |    :|:  | ||   :  :|   | ||     |   |
Db     620 AKSPAEAKSPAEAKSPATVKSPGEAKSPSEAKSPA---EAKSPAEAKSPAEAKSPAEVKS 676

Qy     356 P----APTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPA----PTTTKSAPTTTKEPAPTT 407
           |    :|  || ||    | ||    :  ||   | ||    |   || |   | ||
Db     677 PGEAKSPAEPKSPAEA--KSPA----EVKSPAEAKSPAEVKSPGEAKS-PAAVKSPAEAK 729

Qy     408 TKSAPTTPKE-PSPTTTKEPA----PTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTT 462
           : :|  :| | ||   | ||    |  | |      : | |  || | :  :  :|
Db     730 SPAAVKSPGEAKSPGEAKSPAEAKSPAEAKSPIEVKSPEKAKTPVKEGAKSPAEAKSPEK 789

Qy     463 TKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEP 522
             | |    | ||     | |  :| |     |||  |   : |   || :   | | :
Db     790 AKSPVKEDIKPPAEAKSPEKA-KSPVKEGAKPPEKAKPLDVKSPEAQTPVQEEATVPTDI 848

Qy     523 TPTTPEE-PAPTTPKAAAPNTPKEPAPTT----PKEPAPTTP-------KEPAPTTPKET 570
             |  ||:  :|   || :|   || | |:    ||:    :|       |||     :|
Db     849 RP--PEQVKSPAKEKAKSPE--KEEAKTSEKVAPKKEEVKSPVKEEVKAKEPPKKVEEEK 904

Qy     571 APTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSD--KPAPTTPKGTAPTTP 628
             |||  |  : |: ||    ||  :|   | |::|  :|::  |      |      :
Db     905 TLPTPKTEAKESKKDEAPKEAPKPKVEEKKETPTEKPKDSTAEAKKEEAGEKKKAVASEE 964

Qy     629 KEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTT 688
           : ||    || |   ||    ||  |   |  |:|:     || |: |
Db     965 ETPAKLGVKEEA--KPKEKTETTKTEAEDTKAKEPS----KPTETEKP------------ 1006

Qy     689 PKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSA 748
                 ||  |  |:|                          | :| ||  : |:|
Db    1007 --------KKEEMPAAPEK-----------------KDTKEEKTTESRKPEEK------ 1034

Qy     749 EPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETA 808
             || :|   ||     :| |  :||:     |  :|::     ||      | |:: |
Db    1035 ---PK-MEAKVKEDDKSLSKEP--SKPKTEKAEKSSSTDQKESQPPE------KTTEDKA 1082

Qy     809 TTTEK 813
           |  ||
Db    1083 TKGEK 1087
```

```
RESULT   9
TCNA_TRYCR
ID   TCNA_TRYCR     STANDARD;      PRT;  1162 AA.
AC   P23253;
DT   01-NOV-1991 (Rel. 20, Created)
DT   01-NOV-1991 (Rel. 20, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   SIALIDASE (EC 3.2.1.18) (NEURAMINIDASE) (NA) (MAJOR SURFACE ANTIGEN).
GN   TCNA.
OS   Trypanosoma cruzi.
OC   Eukaryota; Euglenozoa; Kinetoplastida; Trypanosomatidae; Trypanosoma.
OX   NCBI_TaxID=5693;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=SILVIO X-10/4;
RX   MEDLINE=91277609; PubMed=1711561;
RA   Pereira M.E.A., Mejia J.S., Ortega-Barria E., Matzilevich D.,
RA   Prioli R.P.;
RT   "The Trypanosoma cruzi neuraminidase contains sequences similar to
RT   bacterial neuraminidases, YWTD repeats of the low density lipoprotein
RT   receptor, and type III modules of fibronectin.";
RL   J. Exp. Med. 174:179-191(1991).
RN   [2]
RP   SUBCELLULAR LOCATION.
RX   MEDLINE=91376547; PubMed=1896773;
RA   Prioli R.P., Mejia J.S., Aji T., Aikawa M., Pereira M.E.A.;
RT   "Trypanosoma cruzi: localization of neuraminidase on the surface of
RT   trypomastigotes.";
RL   Trop. Med. Parasitol. 42:146-150(1991).
CC   -!- FUNCTION: DEVELOPMENTALLY REGULATED NEURAMINIDASE IMPLICATED IN
CC       PARASITE INVASION OF CELLS.
CC   -!- CATALYTIC ACTIVITY: HYDROLYSIS OF 2,3-, 2,6- AND 2,8-GLYCOSIDIC
CC       LINKAGES JOINING TERMINAL NON-REDUCING N- OR O-ACYLNEURAMINYL
CC       RESIDUES TO GALACTOSE, N-ACETYLHEXOSAMINE, OR N- OR O-ACYLATED
CC       NEURAMINYL RESIDUES IN OLIGOSACCHARIDES, GLYCOPROTEINS,
CC       GLYCOLIPIDS OR COLOMINIC ACID.
```

```
DT   01-MAR-1992 (Rel. 21, Last annotation update)
DE   EXTENSIN PRECURSOR (CELL WALL HYDROXYPROLINE-RICH GLYCOPROTEIN).
GN   HRGPNT3.
OS   Nicotiana tabacum (Common tobacco).
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC   Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots;
OC   Asteridae; euasterids I; Solanales; Solanaceae; Nicotiana.
OX   NCBI_TaxID=4097;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=CV. XANTHI; TISSUE=Leaf;
RX   MEDLINE=90128263; PubMed=2612909;
RA   Keller B., Lamb C.J.;
RT   "Specific expression of a novel cell wall hydroxyproline-rich
RT   glycoprotein gene in lateral root initiation.";
RL   Genes Dev. 3:1639-1646(1989).
CC   -!- FUNCTION: HAS A SPECIALIZED STRUCTURAL FUNCTION, POSSIBLY IN
CC       THE MECHANICAL PENETRATION OF THE CORTEX AND EPIDERMIS OF THE
CC       MAIN ROOT.
CC   -!- SUBCELLULAR LOCATION: EXTRACELLULAR MATRIX.
CC   -!- PTM: EXTENSINS CONTAIN A CHARACTERISTIC REPEAT OF THE PENTAPEPTIDE
CC       SER-PRO(4). THE PROLINE RESIDUE IS HYDROXYLATED AND THEN
CC       GLYCOSYLATED.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X13885; CAA32090.1; -.
DR   PIR; S06733; S06733.
KW   Repeat; Cell wall; Glycoprotein; Signal; Structural protein;
KW   Hydroxylation.
FT   SIGNAL        1      ?
FT   CHAIN         ?    620       EXTENSIN.
FT   REPEAT       70     73        H-A-P-P.
FT   REPEAT      148    151        H-A-P-P.
FT   DOMAIN      229    242        2 X 7 AA TANDEM REPEATS OF T-H-R-H-A-P-P.
FT   REPEAT      229    235        1.
FT   REPEAT      236    242        2.
FT   DOMAIN      205    620        CONTAINS THE SER-PRO(4) REPEATS.
FT   DOMAIN      499    600        3 X APPROXIMATE TANDEM REPEATS.
SQ   SEQUENCE   620 AA;  65406 MW;  641DD2278AB28524 CRC64;

  Query Match             8.1%;  Score 551;  DB 1;  Length 620;
  Best Local Similarity   27.2%;  Pred. No. 3.1e-16;
  Matches  170;  Conservative   64;  Mismatches  304;  Indels   88;  Gaps   24;

Qy     201 AETTTK-GPALTTP--KEPTPTT-----PKEPASTTPKE---PTPTTIKSAPTTPKEPAP 249
           || ||: |  |  |   :| |::     |  | :| |     |:|   : ||     | |
Db      24 AEATTQYGGYLPPPVTSQPPPSSIGLSPPSAPTTTPPSRGHVPSP---RHAPPRHAYPPP 80

Qy     250 TTTKSAPTTPKEP------APTTTKEPAPT---TPKEPAPTTTKEPAPTTTKSAPTTPKE 300
           :     |:    |      |:    | |:   :|  | |:    |         |:  :
Db      81 SHGHLPPSVGGPPPHRGHLPPSRGFNPPPSPVISPSHPPPSYGAPPPSHGPGHLPSHGQR 140

Qy     301 PAPTTPKKPAPT---TPKEPAPTTPKEPTPTTPKEPAPTTK--EPAPTTPKEPAPTAPKK 355
           |       |:|:    |      | |:   | : : |:| ::   | | |  :| ||
Db     141 P-------PSPSHGHAPPSGGHTPPRGQHPPSHRRPSPPSRHGHPPPPTYAQPPPTPIYS 193

Qy     356 PAPTTPKEPAPTTPKEPAPTTTKEPSPTTPK---EPAPTTTKSAPTTTKEPAPTTTKS-- 410
           |:|    :| ||    | | |  :|:|: |    :| | | : || | :   ||   |
Db     194 PSPQV--QPPPT--YSPPPPTHVQPTPSPPSRGHQPQPPTHRHAPPTHRHAPPTHQPSPL 249

Qy     411 ---APTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPT---TPKEPAPTTTK 464
               |:  ::| | |   | |   : | |:    | | |   | |:   :|  || : :
Db     250 RHLPPSPRRQPQPPTYSPPPPAYAQSPQPSPTYSPPPPTYSPPPPSPIYSPPPPAYSPSP 309

Qy     465 KPAPT---APKEPAPTTPKETAPTTPKKL-TPTT------PEKLAPTTPEKPAPTTPEEL 514
```

```
           | ||   :|  || : |   :|  |  |  |::      |   :|  |   :|  |  |
Db     310 PPTPTPTFSPPPPAYSPPPTYSPPPPTYLPLPSSPIYSPPPPVYSPPPPPSYSPPPPTYL 369

Qy     515 APTTPEEPTPTTPEEPAPTTPKAAAP---NTPKEPAPTTPKEPAPTTPKEPAPTTPKETA 571
           |  |  | | :   | ||  ::  |    :|  ||| |   | ||    | | |  :
Db     370 PPPPPSSPPPPSFSPPPPTYEQSPPPPPAYSPPLPAPPTYSPPPPT--YSPPPPTYAQPP 427

Qy     572 PTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEP 631
           |       | |   | |      | |    |  | |      :| |  |  : |     |
Db     428 P------LPPTYSPPPPAYSPPPPPTYSPPPPTYSPPPPAYAQPPPPPPTYSPPPPAYSP 481

Qy     632 APTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKE 691
           | :|    |       | |   | |     | |    |   : || |    |:| |
Db     482 PPPSPIYSPPPPQVQPLPPTFSPPPPRRIHLPPPPHRQP---RPPTPTYGQPPSPPTFSP 538

Qy     692 TAPTTPKEPAPT--TPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAE 749
           |     | |    |: | ||  : | | |     |     |   |: |   ||
Db     539 PPPRQIHSPPPPHWQPRTPTPTYGQPPSPPTFSAPPPRQIHSPPPPHRQPRPPTPTYGQP 598

Qy     750 PTPKALENSPKEPGVPTTKTPAATKP 775
           |:|           ||| :| :  |
Db     599 PSP-----------PTTYSPPSPPP 612
```

```
RESULT   8
NFH_MOUSE
ID   NFH_MOUSE      STANDARD;      PRT;  1087 AA.
AC   P19246; Q61959;
DT   01-NOV-1990 (Rel. 16, Created)
DT   01-AUG-1992 (Rel. 23, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   NEUROFILAMENT TRIPLET H PROTEIN (200 KDA NEUROFILAMENT PROTEIN)
DE   (NEUROFILAMENT HEAVY POLYPEPTIDE) (NF-H).
GN   NEFH OR NFH.
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=89121513; PubMed=3220257;
RA   Julien J.-P., Cote F., Beaudet L., Sidky M., Flavell D., Grosveld F.,
RA   Mushynski W.;
RT   "Sequence and structure of the mouse gene coding for the largest
RT   neurofilament subunit.";
RL   Gene 68:307-314(1988).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=89089138; PubMed=3145094;
RA   Shneidman P.S., Carden M.J., Lees J.F., Lazzarini R.A.;
RT   "The structure of the largest murine neurofilament protein (NF-H) as
RT   revealed by cDNA and genomic sequences.";
RL   Brain Res. 464:217-231(1988).
RN   [3]
RP   SEQUENCE FROM N.A.
RC   STRAIN=SWISS WEBSTER; TISSUE=Brain;
RA   Carden M.J.;
RL   Submitted (MAR-1994) to the EMBL/GenBank/DDBJ databases.
CC   -!- FUNCTION: NEUROFILAMENTS USUALLY CONTAIN THREE IF PROTEINS: L, M,
CC       AND H WHICH ARE INVOLVED IN THE MAINTENANCE OF NEURONAL CALIBER.
CC       NF-H HAS AN IMPORTANT FUNCTION IN MATURE AXONS THAT IS NOT
CC       SUBSERVED BY THE TWO SMALLER NF PROTEINS.
CC   -!- PTM: THERE ARE A NUMBER OF REPEATS OF THE TRIPEPTIDE K-S-P, NFH IS
CC       PHOSPHORYLATED ON A NUMBER OF THE SERINES IN THIS MOTIF. IT IS
CC       THOUGHT THAT PHOSPHORYLATION OF NFH RESULTS IN THE FORMATION OF
CC       INTERFILAMENT CROSS BRIDGES THAT ARE IMPORTANT IN THE MAINTENANCE
CC       OF AXONAL CALIBER.
CC   -!- PTM: PHOSPHORYLATION SEEMS TO PLAY A MAJOR ROLE IN THE FUNCTIONING
CC       OF THE LARGER NEUROFILAMENT POLYPEPTIDES (NF-M AND NF-H), THE
CC       LEVELS OF PHOSPHORYLATION BEING ALTERED DEVELOPMENTALLY AND
CC       COINCIDENT WITH A CHANGE IN THE NEUROFILAMENT FUNCTION.
CC   -!- SIMILARITY: BELONGS TO THE INTERMEDIATE FILAMENT FAMILY.
CC   -!- CAUTION: REF.2 SEQUENCE DIFFERS FROM THAT SHOWN IN POSITIONS 534
```

```
RX   MEDLINE=99018118; PubMed=9799793;
RA   Glockner G., Scherer S., Schattevoy R., Boright A., Weber J.,
RA   Tsui L.C., Rosenthal A.;
RT   "Large-scale sequencing of two regions in human chromosome 7q22:
RT   analysis of 650 kb of genomic sequence around the EPO and CUTL1 loci
RT   reveals 17 genes.";
RL   Genome Res. 8:1060-1073(1998).
RN   [2]
RP   SEQUENCE OF 2338-2700 FROM N.A.
RC   TISSUE=Testis;
RX   MEDLINE=97271566; PubMed=9126492;
RA   Gao Z., Harumi T., Garbers D.L.;
RT   "Chromosome localization of the mouse zonadhesin gene and the human
RT   zonadhesin gene (ZAN).";
RL   Genomics 41:119-122(1997).
CC   -!- FUNCTION: BINDS IN A SPECIES-SPECIFIC MANNER TO THE ZONA PELLUCIDA
CC       OF THE EGG. MAY BE INVOLVED IN GAMETE RECOGNITION AND/OR
CC       SIGNALING.
CC   -!- SUBUNIT: PROBABLY FORMS COVALENT OLIGOMERS.
CC   -!- SUBCELLULAR LOCATION: TYPE I MEMBRANE PROTEIN, EXCLUSIVELY ON THE
CC       APICAL REGION OF THE SPERM HEAD (BY SIMILARITY).
CC   -!- TISSUE SPECIFICITY: IN TESTIS; PRIMARILY IN HAPLOID SPERMATIDS.
CC   -!- DOMAIN: THE MAM DOMAINS PROBABLY MEDIATE SPERM ADHESION TO THE
CC       ZONA PELLUCIDA.
CC   -!- DOMAIN: DURING SPERM MIGRATION THROUGH THE REPRODUCTIVE TRACTS,
CC       THE MUCIN-LIKE DOMAIN MIGHT INHIBIT INAPPROPRIATE TRAPPING OF
CC       SPERMATOZOA OR PROMOTING ADHESION TO THE OVIDUCTAL ISTHMUS.
CC   -!- DOMAIN: THE VWFD DOMAINS 2 AND 3 MAY MEDIATE COVALENT
CC       OLIGOMERIZATION (BY SIMILARITY TO HUMAN INTESTINAL MUCIN MUC2).
CC   -!- SIMILARITY: CONTAINS 3.5 MAM DOMAINS.
CC   -!- SIMILARITY: CONTAINS 4.5 VWFD DOMAINS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; AF053356; AAC78790.1; -.
DR   EMBL; U83191; AAC51208.1; -.
DR   MIM; 602372; -.
DR   InterPro; IPR000561; EGF-like.
DR   InterPro; IPR000998; MAM.
DR   InterPro; IPR002965; P_rich_extensn.
DR   InterPro; IPR002919; TIL.
DR   InterPro; IPR003328; TILa.
DR   InterPro; IPR001007; VWFC.
DR   InterPro; IPR001846; Vwd.
DR   Pfam; PF00629; MAM; 3.
DR   Pfam; PF01826; TIL; 4.
DR   Pfam; PF02345; TILa; 4.
DR   Pfam; PF00094; vwd; 4.
DR   PRINTS; PR01217; PRICHEXTENSN.
DR   SMART; SM00137; MAM; 2.
DR   SMART; SM00214; VWC; 1.
DR   SMART; SM00011; VWC_def; 3.
DR   SMART; SM00216; VWD; 3.
DR   PROSITE; PS01186; EGF_2; 3.
DR   PROSITE; PS00740; MAM_1; 1.
DR   PROSITE; PS50060; MAM_2; 4.
KW   Glycoprotein; Transmembrane; Cell adhesion; Repeat.
FT   NON_TER       1      1
FT   DOMAIN       <1    109       MAM 1.
FT   DOMAIN      112    136       MAM 2 (PARTIAL).
FT   DOMAIN      161    326       MAM 3.
FT   DOMAIN      322    446       MAM 4.
FT   DOMAIN      483    951       66 X HEPTAPEPTIDE REPEATS (APPROXIMATE)
FT                                (MUCIN-LIKE DOMAIN).
FT   DOMAIN      953   1065       VWFD 1 (PARTIAL).
FT   DOMAIN     1066   1454       VWFD 2.
FT   DOMAIN     1455   1861       VWFD 3.
FT   DOMAIN     1862   2292       VWFD 4.
FT   DOMAIN     2293   2684       VWFD 5.
FT   DOMAIN        ?      ?       EGF-LIKE.
FT   CARBOHYD     74     74       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    403    403       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1023   1023       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1099   1099       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1618   1618       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1737   1737       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1832   1832       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1878   1878       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   2136   2136       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   2505   2505       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CONFLICT   2374   2379       NNQKMA -> RAGPGP (IN REF. 1).
FT   NON_TER    2700   2700
SQ   SEQUENCE   2700 AA;  293013 MW;  80E60CC0B12277B1 CRC64;


  Query Match             8.2%;  Score 555.5;  DB 1;  Length 2700;
  Best Local Similarity   32.6%;  Pred. No. 7e-16;
  Matches  199;  Conservative   64;  Mismatches  240;  Indels  107;  Gaps   32;

Qy    123 PINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKT 182
          |:   | ||| | |      |   ||| |: | : |                 | | ||||
Db    445 PVKVLPELPPVSPVSS----TGPSETTGLTENPTISTK-------------KPTVSIEKP 487

Qy    183 SAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPT-----PTTI 237
          |     ||      | |    |: | ::| |   || | |  :   ::||     || :
Db    488 SVTTEKPT-----VPKEKPTIPTEKPTISTEK---PTIPSEKPNMPSEKPTIPSEKPTIL 539

Qy    238 KSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPA-PTTTKSAPT 296
            || | |  ||      ||     |    |:|  |||| |    |  :||  ||   | ||
Db    540 TEKPTIPSE-KPTIPSEKPTISTEKPTVPTEE--PTTPTEETTTYMEEPVIPTEKPSIPT 596

Qy    297 TPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKP 356
            ::|:  | |   ||   |    : ::||    |   || |   || | | :  :|:|
Db    597 --EKPSIPTEK---PTISMEETIISTEKPTICPEKPTIPTEK---PTIPTEKSTISPEK- 647

Qy    357 APTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPK 416
           |||| |  || | |    :|::|:  | | |  :  |    | |   ||     || |
Db    648 -PTTPTE-KPTIPTEKPTISTEKPTIPTEK-PTISPEKLTIPTEKLTIPT---EKPTIPT 701

Qy    417 EPSPTTTKEPAPTTPKEPAPTTPKKPA-----PTTPKEPAPTTPKEPAPTTTKKPAPTAP 471
          |    :|:|  |||| |    : :||:      || | |   |: :|   :| |   |
Db    702 EKPTISTEE--PTTPTEETTISTEKPSIPMEKPTLPTEETTTSVEETTISTEKLTIPM-- 757

Qy    472 KEPAPTTPKETAPT-----TPKKLTPTTPEKLA-----PTTPEKPAPTTPEELAPTTPEE 521
            ::|  :| | | ||     :|:|||  | |||       || | :    : |:|  | | |
Db    758 EKPTISTEKPTIPTEKPTISPEKLTIPT-EKLTIPTEKPTIPIEETTISTEKL--TIPTE 814

Qy    522 PTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPT 581
              :||:|  :| |              || | |  || | |        ||  :| | | ||
Db    815 KPTISPEKPTISTEK-----------PTIPTE-KPTIPTE-------ETTISTEKLTIPT 855

Qy    582 TLKEPAPTTPKKPAPKELAPTTTKEPTSTT------SDKPAPTTPKGTAPTTPKEPAPTT 635
            ::|  :  |   | |    :|:::||  |       ::||   | | | ||  ::
Db    856 --EKPTISPEKLTIPTEKPTISTEKPTIPTEKLTIPTEKPTIPTEKPTIPT--EKLTALR 911

Qy    636 PKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKET--A 693
          |. |:||   | |   :   ||:||       :  |||   :||    | |    |:
Db    912 PPHPSPTA-TGLAALVMSPHAPSTPMTSV---ILGTTTTSRSSTGMSCP-PNARYESCAC 966

Qy    694 PTTPKEPAPT 703
          | : | | |:
Db    967 PASCKSPRPS 976


RESULT   7
EXTN_TOBAC
ID   EXTN_TOBAC     STANDARD;      PRT;   620 AA.
AC   P13983;
DT   01-JAN-1990 (Rel. 13, Created)
DT   01-JAN-1990 (Rel. 13, Last sequence update)
```

```
RESULT   5
FP1_MYTED
ID   FP1_MYTED      STANDARD;      PRT;   875 AA.
AC   Q25460;
DT   30-MAY-2000 (Rel. 39, Created)
DT   30-MAY-2000 (Rel. 39, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ADHESIVE PLAQUE MATRIX PROTEIN (POLYPHENOLIC ADHESIVE PROTEIN) (FOOT
DE   PROTEIN 1) (MEFP1) (FRAGMENT).
GN   FP1.
OS   Mytilus edulis (Blue mussel).
OC   Eukaryota; Metazoa; Mollusca; Bivalvia; Pteriomorphia; Mytiloida;
OC   Mytiloidea; Mytilidae; Mytilus.
OX   NCBI_TaxID=6550;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=91025829; PubMed=1367451;
RA   Filpula D.R., Lee S.M., Link R.P., Strausberg S.L., Strausberg R.L.;
RT   "Structural and functional repetition in a marine mussel adhesive
RT   protein.";
RL   Biotechnol. Prog. 6:171-177(1990).
RN   [2]
RP   PARTIAL SEQUENCE, AND POST-TRANSLATIONAL MODIFICATIONS.
RX   MEDLINE=83135732; PubMed=6298211;
RA   Waite J.H.;
RT   "Evidence for a repeating 3,4-dihydroxyphenylalanine- and
RT   hydroxyproline-containing decapeptide in the adhesive protein of the
RT   mussel, Mytilus edulis L.";
RL   J. Biol. Chem. 258:2911-2915(1983).
CC   -!- FUNCTION: PROVIDES ADHESIVENESS TO THE MUSSEL'S FOOT. MUSSELS
CC       PRODUCE ONE OF THE STRONGEST WATER INSOLUBLE GLUE. THE MUSSEL'S
CC       ADHESIVE IS A BUNDLE OF THREADS, CALLED A BYSSUS, FORMED BY A
CC       FIBROUS COLLAGENOUS CORE COATED WITH ADHESIVE PROTEINS.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- TISSUE SPECIFICITY: PRODUCED BY THE BYSSAL GLAND.
CC   -!- DOMAIN: ALMOST EXCLUSIVELY COMPOSED OF REPEATS OF A DECAPEPTIDE.
CC   -!- PTM: THE DECAPEPTIDE A-K-P-S-Y-P-P-T-Y-K IS POST-TRANSLATIONALLY
CC       MODIFIED AS FOLLOWING: THE SIXTH AND SEVENTH RESIDUES ARE
CC       HYDROXYLATED AND THE PENULTIMATE IS A 3,4-DIHYDROXYPHENYLALANINE
CC       (DOPA) DERIVED FROM TYROSINE.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X54422; CAA38294.1; -.
DR   InterPro; IPR002964; Adhesive_plaq.
DR   InterPro; IPR002965; P_rich_extensn.
DR   PRINTS; PR01216; ADHESIVEI.
DR   PRINTS; PR01217; PRICHEXTENSN.
KW   Repeat; Hydroxylation.
FT   NON_TER       1      1
FT   DOMAIN       67    870       TANDEM REPEATS OF Y-K-[PA]-K-[LP]-[ST]-Y-
FT                                 P-P-[ST].
SQ   SEQUENCE   875 AA;  100412 MW;  6EA85312748CAACE CRC64;


  Query Match             8.2%;  Score 556.5;  DB 1;  Length 875;
  Best Local Similarity   28.7%;  Pred. No. 2.5e-16;
  Matches  272;  Conservative  106;  Mismatches  375;  Indels  195;  Gaps   56;

Qy     107 TQHNKVSTSPKITTAKPINP-----------RP--SLPPNSDTSKETSLTVNKETTVET 152
           |:|  |    | : : |  |             ||  | ||    ||   | : |: :
Db       1 TKHEPVYKPVKTSYSAPYKPPTYQPLKKKVDYRPTKSYPPTYG-SKTNYLPLAKKLSSYK 59

Qy     153 KETTTTNKQTSTDG--KEKTT---SAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKG 207
               || | :|:    | | |   : |   |   |         |:   || ||: :
Db      60 PIKTTYNAKTNYPPVYKPKMTYPPTYKPKPSYPPTYKSKPTYKPKITYPPTYKAKPSY-- 117

Qy     208 PALTTPKEPTPTTPKE----PASTTPKEPTPTTIKSAPTTP-----KEPAPTTTKSAPTT 258
           |:   ||:  | | |      | :  ||   | | |  |: |     |:  |:: |: |:
Db     118 PSSYKPKKTYPPTYKPKLTYPPTYKPKPSYPPTYKPKPSYPPSYKTKKTYPSSYKAKPSY 177

Qy     259 PKEPAPTTTKEPA--PTTPKEPA--------PTTTKEPA-PTTTKSAPTTPK--EPAPTT 305
           |    ||   :|:  ||   :|:         ||   :|  |:| |: |: |    :  ||
Db     178 P----PTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTYPSTYKAKPSYPPTYKAKPTY 233

Qy     306 PKKPA--PTTPKEPA--PTTPKEPT--PTTPKEP---APTTKEPAPTTPKEPA--PTAPK 354
             ||:  ||   :|:  ||   :|:  ||   :|   |  | : ||   :|:  ||
Db     234 KAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTYKAKPTYKAKPSYPPTYKA 293

Qy     355 KPA--PTTPKEPA--PTTPKEPA--PTTTKEPS--PTTPKEPA-PTTTKSAPTTTKEPA- 404
           ||:  ||   :|:  ||   :|:  ||   :||  ||   :|: | | |: ||   :|
Db     294 KPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTY 353

Qy     405 PTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPA--PTTPKEPA--PTTPKEPAP 460
           |:| |: |:  |    ||   :|:     :  ||   ||:  ||   :|:  ||   :|:
Db     354 PSTYKAKPSYP----PTYKAKPSYPPTYKAKPTYKAKPSYPPTYKAKPSYPPTYKAKPSY 409

Qy     461 TTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPE 520
             | |  ||   :|       | |:|  |  |: |    |:    ||:     :  ||
Db     410 PPTYKAKPTYKAKP-------TYPST-YKAKPSYP----PSYKAKPSYPPTYKAKPTYKA 457

Qy     521 EPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKE----PAPTTPKETAPTT-- 574
           :||        |:| ||     |  ||    |   | | |     |:   ||:| | |
Db     458 KPT-------YPSTYKA----KPSYPASYKAKPSYPPTYKSKSSYPSSYKPKKTYPPTYK 506

Qy     575 PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTT----PKE 630
           || |   | | | |: |    ||   | | |                  || : | |    |
Db     507 PKLTYKPTYK-PKPSYPPSYKPKTTYPPTYK-------------PKISYPPTYKAKPSY 551

Qy     631 PAPTTPKEPAPTTPKGT---APTTLKEPA--PTTPKKPA--PKELAPTTTKG-PTSTTSD 682
           ||    |   | | |      ||   :|:  ||   ||:   |   |:| | |:   :
Db     552 PATYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYKAKPTYPSTYKAKPSYPPTY 611

Qy     683 KPAPTTPKETAPTTPKEPA--PTTPKKPA-PTT----PETPPPTTSEVSTPTTTKE---- 731
           |  |: |    ||   :|:  ||   ||  |:|    |  ||     ::| | | |
Db     612 KAKPSYP----PTYKAKPSYPPTYKAKPTYPSTYKAKPSYPPTYKPKISYPPTYKAKPSY 667

Qy     732 PTTIHKSPDESTPELSAEPTPKALENSPK----EPGVPTT--KTPA-----ATKPEMTTT 780
           | | :|:      |   |:|| ||   :|     :|  | |    |:           ||   |
Db     668 PPT-YKAKPSYPPTYKAKPTYKAKPTNPSTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPT 726

Qy     781 AKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKIT 840
             | | |  :   | | |  | |  | :    |   | ::|        :  |    |
Db     727 YKAKPTYKAKPTYPSTYKAKP--TYKAKPTYPPTYKAK-------PSYPPTYKPKPSYPP 777

Qy     841 TLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKK---P 897
           | |: :: |    |||   |    ||: |  |  :        |||  |    |   |
Db     778 TYKSKSIYPSSYKPKKTYPPT---YKPKLTYPPTYK--------PKPSYPPSYKPKITYP 826

Qy     898 TSTKKPKTMPRVRKPKTTPTP-----------RKMTSTMPELNPTSR 933
           :: |   : |   | ||:  |            :  ||  |   ||:|
Db     827 STYKLKPSYPPTYKSKTSYPPTYNKKISYPSSYKAKTSYPPAYKPTNR 874


RESULT   6
ZAN_HUMAN
ID   ZAN_HUMAN      STANDARD;      PRT;  2700 AA.
AC   Q9Y493; O00218;
DT   20-AUG-2001 (Rel. 40, Created)
DT   20-AUG-2001 (Rel. 40, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ZONADHESIN (FRAGMENT).
GN   ZAN.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE OF 1-2379 FROM N.A.
```

```
CC       TUMORS, SUCH AS BREAST CANCER.
CC   -!- PTM: HIGHLY GLYCOSYLATED (N-AND O-LINKED CARBOHYDRATES AND SIALIC
CC       ACID).
CC   -!- POLYMORPHISM: THE NUMBER OF REPEATS IS HIGHLY POLYMORPHIC. IT
CC       VARIES FROM 21 TO 125 IN THE NORTHERN EUROPEAN POPULATION. THE
CC       MOST FREQUENT ALLELES CONTAINS 41 AND 85 REPEATS.
CC   -!- SIMILARITY: CONTAINS 1 SEA DOMAIN.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; J05582; AAA60019.1; -.
DR   EMBL; M32738; AAA35804.1; -.
DR   EMBL; M32739; AAA35806.1; -.
DR   EMBL; J05581; AAA59876.1; -.
DR   EMBL; M61170; AAB53150.1; -.
DR   EMBL; X52229; CAA36478.1; ALT_SEQ.
DR   EMBL; X52228; CAA36477.1; ALT_SEQ.
DR   EMBL; M35093; AAB59612.1; ALT_SEQ.
DR   EMBL; Z17324; CAA78972.1; -.
DR   EMBL; Z17325; CAA78973.1; -.
DR   EMBL; M31823; AAA35757.1; -.
DR   EMBL; S81781; AAD14376.1; ALT_INIT.
DR   EMBL; S81736; AAD14369.1; ALT_INIT.
DR   EMBL; M21868; AAA59874.1; ALT_SEQ.
DR   PIR; A35175; A35175.
DR   PIR; B35175; B35175.
DR   PIR; S10218; S10218.
DR   GlycoSuiteDB; P15941; -.
DR   MIM; 158340; -.
DR   MIM; 113720; -.
DR   InterPro; IPR000082; SEA.
DR   Pfam; PF01390; SEA; 1.
DR   SMART; SM00200; SEA; 1.
DR   PROSITE; PS50024; SEA; 1.
KW   Glycoprotein; Signal; Cytoskeleton; Actin-binding; Transmembrane;
KW   Repeat; Alternative splicing.
FT   SIGNAL        1     23       POTENTIAL.
FT   CHAIN        24   1255       MUCIN 1.
FT   DOMAIN       24   1162       EXTRACELLULAR (POTENTIAL).
FT   TRANSMEM   1163   1186       POTENTIAL.
FT   DOMAIN     1187   1255       CYTOPLASMIC (POTENTIAL).
FT   DOMAIN       81    960       44 X 20 AA TANDEM REPEATS.
FT   DOMAIN     1034   1151       SEA.
FT   CARBOHYD    957    957       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    975    975       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1029   1029       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1055   1055       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1133   1133       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   VARSPLIC     19     19       T -> TATTAPKPAT (IN ISOFORM B).
FT   VARSPLIC     20     22       MISSING (IN ISOFORM C).
FT   VARSPLIC     20     31       MISSING (IN ISOFORM D).
FT   VARSPLIC    126    905       MISSING (IN POLYMORPHIC EPITHELIAL
FT                                ISOFORM).
FT   VARSPLIC   1077   1087       FLQIYKQGGFL -> VSIGLSFPMLP (IN SECRETED
FT                                ISOFORM).
FT   VARSPLIC   1088   1255       MISSING (IN SECRETED ISOFORM).
FT   CONFLICT      2      2       T -> A (IN REF. 11).
FT   CONFLICT    134    134       P -> Q (IN REF. 9).
FT   CONFLICT    154    154       P -> Q (IN REF. 9).
FT   CONFLICT   1021   1021       S -> T (IN REF. 3).
FT   CONFLICT   1251   1251       A -> T (IN REF. 3).
SQ   SEQUENCE   1255 AA;  122072 MW;  5E28DFC4DE7D9A82 CRC64;

 Query Match             9.6%;  Score 651;  DB 1;  Length 1255;
 Best Local Similarity   27.8%;  Pred. No. 5.1e-20;
 Matches  295;  Conservative   72;  Mismatches  460;  Indels  234;  Gaps   51;
```

```
Qy    140 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 198
Db     16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy    199 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTI---K 238
Db     66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVTSAPDNK 125

Qy    239 SAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEP------APTTPKEPAPTTTKEPAPTTTK 292
Db    126 PAPGSTAPPAHGVT-SAPDT--RPAPGSTAPPAHGVTSAPDT--RPAPGSTAPPAHGVT- 179

Qy    293 SAPTTPKEPAPTTP----------KKPAPTTPKEP------APTTPKEPTPTTP------ 330
Db    180 SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT 239

Qy    331 ----KEPAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPK----E 371
Db    240 SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT 299

Qy    372 PAPTTTKEPSPTTPKEPAPTT---TKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEP-- 426
Db    300 SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT-SAPDT--RPAPGSTAPPAH 356

Qy    427 ----APTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEP 474
Db    357 GVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAH 416

Qy    475 APTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTT 534
Db    417 GVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAP----PAHGVTSAPDTRPAPGSTA 472

Qy    535 PKA----AAPNTPKEPAPTTPK----EPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEP 586
Db    473 PPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAP----PAHGVTSAPDTRPAP 528

Qy    587 APTTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPT 634
Db    529 GSTAPPAHGVTSAPDTRPAPGSTAP-PAHGVTSAPDTRPAPGSTAPPAHGVTS----APD 583

Qy    635 TPKEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTT 680
Db    584 TRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAP-PAHGVTSAP 642

Qy    681 SDKPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVS 724
Db    643 DTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGS--TAPPAHGVTS 700

Qy    725 TPTTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEM 777
Db    701 APDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAP-----DTRPAPGSTAP-- 753

Qy    778 TTTAKDKTTERDLRTTPETT-------TAAPKMTKETATTTEKTTESKITATTTQVTSTT 830
Db    754 --PAHGVTSAPDTRPAPGSTAPPAHGVTSAP----DTRPAPGSTAPPAHGVTSAPDTRPA 807

Qy    831 TQDTTP--FKITTLKTTTLAPKVTTTKKTITTT--EIMNKPEETAKPKDRATNSKATTPK 886
Db    808 PGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPA 867

Qy    887 PQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPN 946
Db    868 P--GSTAPPAHGVTSAPDTRP---APGSTAPPAHGVTSAPDTRPA----------PG 909

Qy    947 QT-PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPML 1005
Db    910 STAPPAHGVTSAPDTRPAPGSTAPPAH----------GVTSAPDNRPALGSTA----PPV 955

Qy   1006 SDETNICNGKPVDGLTTLRNGTLVAFRGHYFWMLSPFSPPS 1046
Db    956 HNVTSASGSASGSASTLVHNGTSARATTTPASKSTPFSIPS 996
```

```
           ||  |  : ::|        |  |:     ||  |||      |     ||   :   :|
Db    905 YPGSQTETSVSSTTETTIVPTKTTTSVTTPSTTTITTTVCSTGTNSAGETTSGCSPKTVT 964

Qy    821 AT--TTQVTSTTTQDTTPFKITTLKTTTLAPKVTT---TKKTITTT-EIMNKPEETAKPK 874
          |  ||  || ||  ||     |   | | :   ||   : ||||||      | |||
Db    965 TTVPTTTTTSVTTSSTTTITTTVCSTGTNSAGETTSGCSPKTITTTVPCSTSPSETA--- 1021

Qy    875 DRATNSKATTPKPQKP----------------TKAPKKPTSTKKPKTMPRVRKPKTTPT 917
            |::||  |  |                    ||   : |:|   | :|        ||
Db   1022 -----SESTTTSPTTPVTTVVSTTVVTTEYSTSTKPGGEITTTFVTKNIPTTYLTTIAPT 1076

Qy    918 PRKMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRP 977
          |   :|:    ||: |   :  | |             |   |   |  :|    :
Db   1077 P--SVTTVTNFTPTT-ITTTVCSTGT-----------NSAGETTSGCSPKTVTTTVPC 1120

Qy    978 HVFMPEVTPDMDYLPRVPNQGIIINPMLSDETNICNGKPVDGLTTLRNGTLVAFRGHYFW 1037
            | | :   |        ::    |  ||   ||   | ||    |       |
Db   1121 STGTGEYTTEATTLVTTAVTTTVVTTESSTGTNSA-GKTTTGYTTKSVPTT------YVT 1173

Qy    1038 MLSPFSPPSPARRITEVWGIPSPIDTVFTRC----NCEGKT 1074
             |:| :| :||       :|: | |  | |    |  |:|
Db    1174 TLAPSAPVTPATN-----AVPTTITT--TECSAATNAAGET 1207
```

```
RESULT   4
MUC1_HUMAN
ID   MUC1_HUMAN     STANDARD;      PRT;  1255 AA.
AC   P15941; P15942; P13931; P17626; Q14128; Q16442; Q16437; Q9Y4J2;
DT   01-JAN-1990 (Rel. 13, Created)
DT   01-APR-1990 (Rel. 14, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   MUCIN 1 PRECURSOR (POLYMORPHIC EPITHELIAL MUCIN) (PEM) (PEMT)
DE   (EPISIALIN) (TUMOR-ASSOCIATED MUCIN) (CARCINOMA-ASSOCIATED MUCIN)
DE   (TUMOR-ASSOCIATED EPITHELIAL MEMBRANE ANTIGEN) (EMA) (H23AG) (PEANUT-
DE   REACTIVE URINARY MUCIN) (PUM) (BREAST CARCINOMA-ASSOCIATED ANTIGEN
DE   DF3).
GN   MUC1.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Pancreas;
RX   MEDLINE=90368716; PubMed=2394722;
RA   Lan M.S., Batra S.K., Qi W.-N., Metzgar R.S., Hollingsworth M.A.;
RT   "Cloning and sequencing of a human pancreatic tumor mucin cDNA.";
RL   J. Biol. Chem. 265:15294-15299(1990).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=90202794; PubMed=2318825;
RA   Ligtenberg M.J.L., Vos H.L., Gennissen A.M.C., Hilkens J.;
RT   "Episialin, a carcinoma-associated mucin, is generated by a
RT   polymorphic gene encoding splice variants with alternative amino
RT   termini.";
RL   J. Biol. Chem. 265:5573-5578(1990).
RN   [3]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Breast carcinoma;
RX   MEDLINE=90368715; PubMed=1697589;
RA   Gendler S.J., Lancaster C.A., Taylor-Papadimitriou J., Duhig T.,
RA   Peat N., Burchell J., Pemberton L., Lalani E.-N., Wilson D.;
RT   "Molecular cloning and expression of human tumor-associated
RT   polymorphic epithelial mucin.";
RL   J. Biol. Chem. 265:15286-15293(1990).
RN   [4]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=91097524; PubMed=2268309;
RA   Lancaster C.A., Peat N., Duhig T., Wilson D.,
RA   Taylor-Papadimitriou J., Gendler S.J.;
RT   "Structure and expression of the human polymorphic epithelial mucin
RT   gene: an expressed VNTR unit.";
RL   Biochem. Biophys. Res. Commun. 173:1019-1029(1990).
RN   [5]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Breast carcinoma;
RX   MEDLINE=90276413; PubMed=2351132;
RA   Wreschener D.H., Hareuveni M., Tsarfaty I., Smorodinsky N.,
RA   Horov J., Zaretsky J., Kotkes P., Weiss M., Lathe R., Dion A.,
RA   Keydar I.;
RT   "Human epithelial tumor antigen cDNA sequences. Differential splicing
RT   may generate multiple protein forms.";
RL   Eur. J. Biochem. 189:463-473(1990).
RN   [6]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Breast carcinoma;
RX   MEDLINE=90276414; PubMed=2112460;
RA   Hareuveni M., Tsarfaty I., Zaretsky J., Kotkes P., Horev J.,
RA   Zrihan S., Weiss M., Green S., Lathe R., Keydar I., Wreschner D.H.;
RT   "A transcribed gene, containing a variable number of tandem repeats,
RT   codes for a human epithelial tumor antigen. cDNA cloning, expression
RT   of the transfected gene and over-expression in breast cancer
RT   tissue.";
RL   Eur. J. Biochem. 189:475-486(1990).
RN   [7]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=91033045; PubMed=1688329;
RA   Tsarfaty I., Hareuveni M., Horev J., Zaretsky J., Weiss M.,
RA   Jeltsch J.M., Garnier J.M., Lathe R., Keydar I., Wreschner D.H.;
RT   "Isolation and characterization of an expressed hypervariable gene
RT   coding for a breast-cancer-associated antigen.";
RL   Gene 93:313-318(1990).
RN   [8]
RP   PARTIAL SEQUENCE FROM N.A.
RX   MEDLINE=88330762; PubMed=3417635;
RA   Gendler S.J., Taylor-Papadimitriou J., Duhig T., Rothbard J.,
RA   Burchell J.;
RT   "A highly immunogenic region of a human polymorphic epithelial mucin
RT   expressed by carcinomas is made up of tandem repeats.";
RL   J. Biol. Chem. 263:12820-12823(1988).
RN   [9]
RP   SEQUENCE OF 1-169 FROM N.A.
RX   MEDLINE=90088473; PubMed=2597151;
RA   Abe M., Siddiqui J., Kufe D.;
RT   "Sequence analysis of the 5' region of the human DF3 breast
RT   carcinoma-associated antigen gene.";
RL   Biochem. Biophys. Res. Commun. 165:644-649(1989).
RN   [10]
RP   SEQUENCE OF 1-109 FROM N.A.
RC   TISSUE=Thyroid;
RX   MEDLINE=96183746; PubMed=8608966;
RA   Weiss M., Baruch A., Keydar I., Wreschner D.H.;
RT   "Preoperative diagnosis of thyroid papillary carcinoma by reverse
RT   transcriptase polymerase chain reaction of the MUC1 gene.";
RL   Int. J. Cancer 66:55-59(1996).
RN   [11]
RP   SEQUENCE OF 1-89 FROM N.A.
RC   TISSUE=Lung;
RX   MEDLINE=96181716; PubMed=8604237;
RA   Yu C.J., Yang P.C., Shew J.Y., Hong T.M., Yang S.C., Lee Y.C.,
RA   Lee L.N., Luh K.T., Wu C.W.;
RT   "Mucin mRNA expression in lung adenocarcinoma cell lines and
RT   tissues.";
RL   Oncology 53:118-126(1996).
RN   [12]
RP   SEQUENCE OF 1-46 FROM N.A.
RC   TISSUE=Breast carcinoma;
RA   Buluwela L., Liu Q., Luqmani Y.A., Gomm J.J., Coombes R.C.;
RL   Submitted (OCT-1992) to the EMBL/GenBank/DDBJ databases.
CC   -!- FUNCTION: DIRECT OR INDIRECT INTERACTION WITH ACTIN
CC       CYTOSKELETON.
CC   -!- SUBCELLULAR LOCATION: TYPE I MEMBRANE PROTEIN. A SECRETED FORM
CC       IS ALSO PRODUCED.
CC   -!- ALTERNATIVE PRODUCTS: VARIOUS VARIANTS ARE PRODUCED BY ALTERNATIVE
CC       SPLICING.
CC   -!- TISSUE SPECIFICITY: ABERRANTLY EXPRESSED IN HUMAN EPITHELIAL
```

```
Qy     922 -------TSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHM- 973
                  |  || | |   |  :             :||:     :::: ||:  :|:
Db    1454 YLRGYPDGSFRPERNITRAEAAVIF----------AKLL----GADESYGAQSASPYSD 1498

Qy     974 LLRPH---------------------VFMPE--------VTPDMDYLPRVPNQGI----- 999
            |   |                     | |:          |  : :| :|  | |
Db    1499 LADTHWAAWAIKFATSQGLFKGYPDGTFKPDQNITRAEFATVVLHFLTKVKGQEIMSKLA 1558

Qy    1000 ---IINPMLSDETNICNG----KPVDGLTTL------RNGTLVAFRGHYFWMLSPFSPPS 1046
             | ||   |    | |    : :: ||:|       :||               | | :
Db    1559 TIDISNPKFDD----CVGHWAQEFIEKLTSLGYISGYPDGT-------------FKPQN 1600

Qy    1047 PARRITEVWGIPSPIDTVFTRCNCEGKTFFFKD--SQYWRFTNDIKD 1091
             :|   |    : |:    |    |    | |    || |  || |
Db    1601 YIKRSESV----ALINRALERGPLNGAPKLFPDVNESYWAF-GDIMD 1642
```

```
RESULT   3
AMYH_YEAST
ID   AMYH_YEAST     STANDARD;      PRT;  1367 AA.
AC   P08640; P08068;
DT   01-AUG-1988 (Rel. 08, Created)
DT   01-FEB-1995 (Rel. 31, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   GLUCOAMYLASE S1/S2 PRECURSOR (EC 3.2.1.3) (GLUCAN 1,4-ALPHA-
DE   GLUCOSIDASE) (1,4-ALPHA-D-GLUCAN GLUCOHYDROLASE).
GN   STA1 OR STA2 OR MAL5 OR YIR019C.
OS   Saccharomyces cerevisiae (Baker's yeast).
OC   Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
OC   Saccharomycetales; Saccharomycetaceae; Saccharomyces.
OX   NCBI_TaxID=4932;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=S288C / AB972;
RA   Barrell B.G., Badcock K., Bankier A.T., Bowman S., Brown D.,
RA   Churcher C.M., Connor R., Copsey T., Dear S., Devlin K., Fraser A.,
RA   Gentles S., Hamlyn N., Horsnell T.S., Hunt S., Jagels K., Jones M.,
RA   Louis E., Lye G., Moule S., Moule T., Odell C., Pearson D.,
RA   Rajandream M.A., Riles L., Rowley N., Skelton J., Smith V.,
RA   Walsh S.V., Whitehead S.;
RL   Submitted (DEC-1994) to the EMBL/GenBank/DDBJ databases.
RN   [2]
RP   SEQUENCE OF 1-242 AND 762-1331 FROM N.A.
RX   MEDLINE=87194600; PubMed=3106330;
RA   Yamashita I., Nakamura M., Fukui S.;
RT   "Gene fusion is a possible mechanism underlying the evolution of
RT   STA1.";
RL   J. Bacteriol. 169:2142-2149(1987).
RN   [3]
RP   SEQUENCE OF 1-31 FROM N.A.
RC   STRAIN=SPX101-1C;
RX   MEDLINE=89031230; PubMed=3141213;
RA   Pardo J.M., Ianez E., Zalacain M., Claros M.G., Jimenez A.;
RT   "Similar short elements in the 5' regions of the STA2 and SGA genes
RT   from Saccharomyces cerevisiae.";
RL   FEBS Lett. 239:179-184(1988).
CC   -!- CATALYTIC ACTIVITY: HYDROLYSIS OF TERMINAL 1,4-LINKED ALPHA-D-
CC       GLUCOSE RESIDUES SUCCESSIVELY FROM NON-REDUCING ENDS OF THE CHAINS
CC       WITH RELEASE OF BETA-D-GLUCOSE.
CC   -!- SIMILARITY: TO S.POMBE SPBC215.13.
CC   -!- SIMILARITY: SOME, TO S.POMBE SPCC285.13C.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; Z38061; CAA86176.1; -.
DR   EMBL; M16164; AAA35014.1; -.
DR   EMBL; M16165; AAA35015.1; -.
DR   EMBL; X13857; CAA32069.1; -.
DR   PIR; B26877; B26877.
DR   PIR; A26877; A26877.
DR   PIR; S48478; S48478.
DR   SGD; S0001458; MUC1.
KW   Hydrolase; Glycosidase; Polysaccharide degradation; Glycoprotein;
KW   Signal; Multigene family.
FT   SIGNAL       1     21       POTENTIAL.
FT   CHAIN       22   1367       GLUCOAMYLASE S1/S2.
FT   DOMAIN     210   1367       SER/THR-RICH.
FT   CARBOHYD   817    817       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   874    874       N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE   1367 AA;  136110 MW;  91C00E2DBD61AA9D CRC64;
```

```
  Query Match             11.7%;  Score 798.5;  DB 1;  Length 1367;
  Best Local Similarity   27.9%;  Pred. No. 5.9e-26;
  Matches  313;  Conservative  108;  Mismatches  523;  Indels  177;  Gaps   50;

Qy      44 DCDAQCKKYDK---CCPDYESFCAEVK-DNKKNRTKKKPTPKPPVVDE--AGSGLDNGDF 97
           |    |  ||       |:  |   :  ||    ||    |          |  ::
Db     174 DLSTGCNNYDNQGHSQTDFPGFYWNIDCDNNCGGTKSSTTTSSTSESSTTTSSTSESSTT 233

Qy      98 KVTTPDTSTTQHNKVSTSPKITTAKPINP--------RPSLPPNSDTSKETSLTVNKETT 149
             :| ::|||  :   :|    :|  |  |          :|: |  :  :||     : :||
Db     234 TSSTSESSTTTSSTSESSTSSSTTAPATPTTTSCTKEKPTPPTTTSCTKEKPTPPHHDTT 293

Qy     150 VETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETT--TKG 207
             ||: |||:| | |   :|||:   | |   |:    ||       ||| : ||  :
Db     294 PCTKKKTTTSK-TCT---KKTTTPVPTPS-SSTTESSSAPV------PTPSSSTTESSSA 342

Qy     208 PALTTPKE------PTPTTPKEPASTTPKEPTPTTIKSAP----TTPKEPAPTTTKSAPT 257
           |  ::  |       |||::    :|: |   : |   |||    ||    ||  | |: |
Db     343 PVTSSTTESSSAPVPTPSSSTTESSSAPVTSSTTESSSAPVTSSTTESSSAPVPTPSSST 402

Qy     258 TPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEP 317
           |    || |:     ||    || |:   : | : ||| |    : ||    || |
Db     403 TESSSAPVTSS----TTESSSAPVTS---STTESSSAPVT----SSTTESSSAPVTSSTT 451

Qy     318 APTTPKEPTP---TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAP----TTPK 370
             ::   |||   ||    || |  : ||    ||  |   : ||    ||   ||
Db     452 ESSSAPVPTPSSSTTESSSAPVT---SSTTESSSAP-VPTPSSSTTESSSAPVTSSTTES 507

Qy     371 EPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAP----TTTKSAPTTPKEPSPTTTKEP 426
             ||  |   | ||    ||  | |: ||    ||    ||  |:   |  | ||
Db     508 SSAPVPTPS-SSTTESSSAPAPTPSSSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSS 566

Qy     427 AP----TTPKEPAPT-TPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKE 481
           |    ||    ||  ||       :   | ||    |: :||:  :  ||    : ||
Db     567 TPVTSSTTESSSAPVPTPSSSTTESSSAPVPT----PSSSTTESSSAPAPTPSSSTTESS 622

Qy     482 TAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPT-TPEEPTPTTPEEPAP-----TTP 535
           :|| |    : ||    ||  |   : ||    ||  ||    |   :   | |    ||
Db     623 SAPVT----SSTTESSSAP-VPTPSSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTE 677

Qy     536 KAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPT-TPKGTAPTTLKEPAPTTPKKP 594
           ::|| |    : ||    || |    : ||   :||  ||  :   :  | ||
Db     678 SSSAPVT----SSTTESSSAPVT----SSTTESSSAPVPTPSSSTTESSSAPVPTPSSST 729

Qy     595 APKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAP---TTPKEPAPT-TPKGTAPT 650
             ||  |   :|||    || |   |  ::   | |   ||    ||  ||  :
Db     730 TESSSAPVPTPS-SSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTE 788

Qy     651 TLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAP-----TTP 705
           :   | ||         :||  |   :|  :      :||  ::  :   | |     ||
Db     789 SSSAPVPTPSSSTTESSVAPVPTPSSSSNITSSAPSSTPFSSSTESSSVPVPTPSSSTTE 848

Qy     706 KKPAPTTPETPPPTTSEVSTPT-----TTKEPTTIHKSPDESTPELSAEPTPKALENSPK 760
             || :  |   : : | ||:     |:  |::|   |  || | :  |      :| |
Db     849 SSSAPVSSSTTESSVAPVPTPSSSSNITSSAPSSI---PFSSTTESFSTGT-TVTPSSSK 904

Qy     761 EPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKIT 820
```

```
Qy     565 T-TPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAP-TTPKG 622
           | |  :|  |||  |  |    | ||  : |    :  |||  || | :    | |||
Db    1958 TPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPIT 2017

Qy     623 TAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSD 682
           |  |    | ||  : |  |||  |  |    | ||  : |    :  |||  || | :
Db    2018 TTTTVTPTPTPTGTQTPT-TTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTG 2076

Qy     683 KPAP-TTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDE 741
             | |||  |  |    | ||  : |  |||  |   | :   ||| |: |||   :|
Db    2077 TQTPTTTPITTTTTVTPTPTPTGTQTPT-TTPITTTTTVTPTPTPTGTQTPTT---TPIT 2132

Qy     742 STPELSAEPTPKALENSPKEP----------GVPT-TKTPAATKPEMTTTAKDKTTERDL 790
           :|  ::  |||   :     |              || |:||  |    |||      |
Db    2133 TTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGT 2192

Qy     791 R---TTP--ETTTAAPKMT---KETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTL 842
           :   |||   |||  |  |    :| |||  || : :| | |  | | |  |||  |||
Db    2193 QTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTP-TGTQTPTTTP--ITT- 2248

Qy     843 KTTTLAPKVTTT-KKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAP-KKPTST 900
            |||: |  | |  :| ||| |      |  |    | :  |||      |  |   || |
Db    2249 -TTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGT 2307

Qy     901 KKPKTMPRVRKPKTTPTPRKMTSTMPELNP---TSRIAEAMLQTTTR-PNQTPNSKLVEV 956
           : | | |       ||||    :  |   |   |: :       | |: | || :    |
Db    2308 QTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTV 2367

Qy     957 NPKSEDAGGAEGETPHMLLRPHVFMPEVTP 986
           |    |    :||      |     |||
Db    2368 TPTPTPTG---TQTP--TTTPITTTTTVTP 2392
```

```
RESULT   2
SLP1_CLOTM
ID   SLP1_CLOTM     STANDARD;      PRT;  1664 AA.
AC   Q06852;
DT   01-JUN-1994 (Rel. 29, Created)
DT   01-JUN-1994 (Rel. 29, Last sequence update)
DT   01-FEB-1996 (Rel. 33, Last annotation update)
DE   CELL SURFACE GLYCOPROTEIN 1 PRECURSOR (OUTER LAYER PROTEIN B) (S-LAYER
DE   PROTEIN 1).
GN   OLPB.
OS   Clostridium thermocellum.
OC   Bacteria; Firmicutes; Bacillus/Clostridium group; Clostridiaceae;
OC   Clostridium.
OX   NCBI_TaxID=1515;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=NCIB 10682;
RX   MEDLINE=93209931; PubMed=8458832;
RA   Fujino T., Beguin P., Aubert J.-P.;
RT   "Organization of a Clostridium thermocellum gene cluster encoding the
RT   cellulosomal scaffolding protein CipA and a protein possibly involved
RT   in attachment of the cellulosome to the cell surface.";
RL   J. Bacteriol. 175:1891-1899(1993).
CC   -!- SUBUNIT: ASSEMBLED INTO MONO-LAYERED CRYSTALLINE ARRAYS.
CC   -!- SUBCELLULAR LOCATION: CELL WALL.
CC   -!- SIMILARITY: CONTAINS 4 S-LAYER HOMOLOGY (SLH) DOMAINS.
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; X67506; CAA47841.1; -.
DR   InterPro; IPR001119; SLH.
DR   Pfam; PF00395; SLH; 3.
DR   PROSITE; PS01072; SLH_DOMAIN; 2.
KW   Cell wall; S-layer; Signal; Repeat.
FT   SIGNAL       1     28       POTENTIAL.
FT   CHAIN       29   1664       CELL SURFACE GLYCOPROTEIN 1.
FT   DOMAIN      36    763       4 X 156 AA APPROXIMATE REPEATS.
FT   REPEAT      36    191       1.
FT   REPEAT     207    363       2.
FT   REPEAT     409    565       3.
FT   REPEAT     607    763       4.
FT   DOMAIN     771   1377       APPROXIMATE TANDEM REPEATS OF
FT                               T-P-S-D-E-P.
FT   DOMAIN    1378   1449       GLY/PRO/SER/THR-RICH.
FT   DOMAIN    1453   1494       SLH 1 (INCOMPLETE).
FT   DOMAIN    1495   1565       SLH 2.
FT   DOMAIN    1566   1625       SLH 3.
FT   DOMAIN    1626   1646       SLH 4 (INCOMPLETE).
SQ   SEQUENCE   1664 AA;  178194 MW;  5F396695BA9FE74B CRC64;
```

```
  Query Match             13.9%;  Score 950;  DB 1;  Length 1664;
  Best Local Similarity   31.0%;  Pred. No. 5.1e-32;
  Matches  312;  Conservative  104;  Mismatches  361;  Indels  230;  Gaps   55;

Qy     193 VLAKPTP-KAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTT 251
           |: :| | || :    |  |   ||||:          ||||:     ||   || |:
Db     758 VVIQPAPIKAASDEPIPTDTPSDEPTPS----------DEPTPS---DEPTPSDEPTPSD 804

Qy     252 TKSAPTTPKEPAPTTTKEPAPTTPKEPA----PTTTKEPAPTTTKSAPTTPKEPAPTTPK 307
             :  ||:|| || |    ||   ||      || : || |:   :   ||:|| ||
Db     805 EPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTP 864

Qy     308 KPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPT-TPKEPAPTAPKKPAPTTPKEPAP 366
             ||   || |:   ||||:   || | : || |: ||:|| ||      ||   || |
Db     865 SDEPTPSDEPTPS--DEPTPS--DEPTP-SDEPTPSETPEEPIPTDTPSDEPTPSDEPTP 919

Qy     367 TTPKEPA--PTTTKEPSPT-TPKEPAPTTTKS-APTTTKEPAPT---TTKSAPTTPKEPS 419
           :    |:  || : ||:|: ||:|| || | |  || : || |:   |    ||   ||:
Db     920 SDEPTPSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPT 979

Qy     420 ----PTTTKEPAPT-TPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEP 474
               || : || |: ||:|| ||      ||   || |:    |:   |    ||   ||
Db     980 PSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSDEP 1039

Qy     475 APT---TPKETAPTTPKKLTPTTPEKLAPTTPKEPAPT---TPEELAPTTPEEPTPT-TP 527
            |:   || |   ||::  ||      ||  ::| |:   || :  ||  :||||: ||
Db    1040 TPSDEPTPSE----TPEEPIPTDTPSDEPTPSDEPTPSDEPTPSD-EPTPSDEPTPSETP 1094

Qy     528 EEPAPTTPKAAAPNTPKEPAPTTPKEPA--PTTPKEPAPT-TPKETAPTTPKGTAPTTLK 584
           ||| ||   :  |    || |:    |:  ||   || |: ||:|  ||      ||
Db    1095 EEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSD 1154

Qy     585 EPAPTTPKKPAPKELAPTTTKEPT-STTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTT 643
           || |:   :| | :  || : ||| | | ::| ||      ||   || |:   || |:
Db    1155 EPTPS--DEPTPSD-EPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPS--DEPTPS- 1208

Qy     644 PKGTAPTTLKEPAPT-TPKKPAPKELAPTTTKGPTSTTSDKPAPT---TPKETAPTTPKE 699
             ||   || |: ||::|           || | ||:| |:   || :  ||   |
Db    1209 ---DEPTPSDEPTPSETPEEPI----------PTDTPSDEPTPSDEPTPSD-EPTPSDE 1253

Qy     700 PAPTTPKKPAPT-TPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENS 758
           | |:   :| |: ||| | || :    || : |||      || ||  | ||||
Db    1254 PTPS--DEPTPSETPEEPIPTDTPSDEPTPSDEPT----PSDEPTP--SDEPTPSDEPTP 1305

Qy     759 PKEPGVPTTKTPAAT--KPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTE 816
             ||      ||: |  :|  | |  |: |  |      | :  |  : |   : | |
Db    1306 SDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSD----EPTPSDEPTPSDEPTPSDEPTPS 1361

Qy     817 SKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDR 876
           : | :     | | :: ||       |||  |  :||                      |
Db    1362 DEPTPSDEPTPSETPEEPTP-------TTTPTPTPSTT-----------------PTSG 1396

Qy     877 ATNSKAT--------TPKPQKPTKAP-KKPTSTKKPKTMPRVRKPKTTPTPRKM------ 921
           :  |  :           |  ||  |  |||||  |     : :|  :  |  :
Db    1397 SGGSGGSGGGGGGGGGTVPTSPTPTPTSKPTSTPAP---TEIEEPTPSDVPGAIGGEHRA 1453
```

```
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------------
DR   EMBL; L21998; AAB95295.1; -.
DR   EMBL; M74027; AAA59875.1; -.
DR   EMBL; M94131; AAA59163.1; -.
DR   EMBL; M94132; AAA59164.1; -.
DR   MIM; 158370; -.
DR   InterPro; IPR000359; Cys_knot.
DR   InterPro; IPR000561; EGF-like.
DR   InterPro; IPR002400; GF_cysknot.
DR   InterPro; IPR001007; VWFC.
DR   InterPro; IPR001846; Vwd.
DR   Pfam; PF00007; Cys_knot; 1.
DR   Pfam; PF00094; vwd; 4.
DR   PRINTS; PR00438; GFCYSKNOT.
DR   SMART; SM00214; VWC; 2.
DR   SMART; SM00011; VWC_def; 2.
DR   SMART; SM00216; VWD; 4.
DR   PROSITE; PS00022; EGF_1; UNKNOWN_1.
DR   PROSITE; PS01185; CTCK_1; 1.
DR   PROSITE; PS01225; CTCK_2; 1.
DR   PROSITE; PS01208; VWFC; 2.
KW   Glycoprotein; Repeat; Signal.
FT   SIGNAL        1     20       POTENTIAL.
FT   CHAIN        21   5179       MUCIN 2.
FT   DOMAIN     1401   1747       APPROXIMATE REPEATS.
FT   REPEAT     1401   1416       1.
FT   REPEAT     1417   1432       2.
FT   REPEAT     1433   1448       3.
FT   REPEAT     1449   1464       4.
FT   REPEAT     1465   1471       5.
FT   REPEAT     1472   1478       6.
FT   REPEAT     1479   1494       7A.
FT   REPEAT     1495   1517       7B.
FT   REPEAT     1518   1533       8A.
FT   REPEAT     1534   1556       8B.
FT   REPEAT     1557   1572       9A.
FT   REPEAT     1573   1596       9B.
FT   REPEAT     1597   1612       10A.
FT   REPEAT     1613   1635       10B.
FT   REPEAT     1636   1651       11A.
FT   REPEAT     1652   1675       11B.
FT   REPEAT     1676   1683       12.
FT   REPEAT     1684   1699       13.
FT   REPEAT     1700   1715       14.
FT   REPEAT     1716   1731       15.
FT   REPEAT     1732   1747       16.
FT   DOMAIN     4815   4886       VWFC 1.
FT   DOMAIN     4924   4991       VWFC 2.
FT   DOMAIN     5075   5160       CTCK.
FT   DISULFID   5075   5122       BY SIMILARITY.
FT   DISULFID   5089   5136       BY SIMILARITY.
FT   DISULFID   5098   5152       BY SIMILARITY.
FT   DISULFID   5102   5154       BY SIMILARITY.
FT   DISULFID      ?   5159       BY SIMILARITY.
FT   CARBOHYD    163    163       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    423    423       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    670    670       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    770    770       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    894    894       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1139   1139       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1154   1154       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1215   1215       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1230   1230       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1246   1246       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1787   1787       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1820   1820       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4339   4339       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4351   4351       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4362   4362       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4373   4373       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4422   4422       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4438   4438       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4502   4502       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4616   4616       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4627   4627       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4752   4752       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4787   4787       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4881   4881       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4888   4888       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4955   4955       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4970   4970       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   5019   5019       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   5038   5038       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   5069   5069       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CONFLICT   1351   1351       H -> L (IN REF. 3).
FT   CONFLICT   1412   1412       T -> S (IN REF. 3).
FT   CONFLICT   1449   1449       L -> P (IN REF. 3).
FT   CONFLICT   1504   1504       M -> T (IN REF. 3).
FT   CONFLICT   4192   4192       G -> S (IN REF. 2).
SQ   SEQUENCE   5179 AA;  540295 MW;  85CD7571FB9A5663 CRC64;

  Query Match             17.2%;  Score 1171;  DB 1;  Length 5179;
  Best Local Similarity   32.8%;  Pred. No. 1.5e-40;
  Matches  364;  Conservative   57;  Mismatches  461;  Indels  228;  Gaps   41;

Qy      38 ERGRECDCDAQ----CKKYDK-------CCPDYE--SFCAEVKDNKKNRTKKKPTPKPPV 84
           : |::  ||       ||  |:        | ||:    |    |      |   || ||
Db    1350 QHGQKVQCDVSVGFICKNEDQFGNGPFGLCYDYKIRVNCCWPMD------KCITTPSPPT 1403

Qy      85 VDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTSKE--TSL 142
                      : | |:||      | |  ||  |     : || : |     |:
Db    1404 TTP-----------SPPPTTTT------TLPPTTTPSPPTTTTTTPPPTTTPSPPITTT 1445

Qy     143 TVNKETTVETK--ETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPK 200
           |    ||  :   ||||   |:|     || :  | :                    |:|
Db    1446 TTPLPTTTPSPPISTTTTPPPTTTPSPPTTTPSPPTTT----------------PSPP 1487

Qy     201 AETTTKGPALTTPKEP--TPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTT 258
           |||  |  |||  |  || ||  |||||   || ||     ||   | |||| | |||
Db    1488 TTTTTTPPPTTTPSPPMTTPITP--PASTTTLPPT-TTPSPPTTTTTTPPPTTTPSPPTT 1544

Qy     259 PKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEPA 318
             ||:|    |||   | ||||  | |||| | |||      |||  |  ||   |
Db    1545 TPITPPTSTTTLPPTTTPSPPPTTTTTPPPTTTPSPPTT------TTPSPPTITTTTPPP 1598

Qy     319 PTTPKEPTPTTPKEPAPTTKEPAPTTPKEP-------APTAPKKPAPTTPKEPAPTTPKE 371
           |||  || ||   |  ||  |  |||  |         ||    | |||   | |||
Db    1599 TTTPSPPTTTTTTPPPTTTPSPPTTTPITPPTSTTTLPPTTTPSPPPTTTTTPPPTTTPS 1658

Qy     372 PAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAP---TTPKEPSPTTTKEPAP 428
           |  |||  |  ||   | |||| |:| ||    |||| : |   |||  |  |||  |:
Db    1659 PPTTTTPSPPITTTTTPPPTTTPSSPITTTPSPPTTTMTTPSPTTTPSSPI-TTTTTPSS 1717

Qy     429 TTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTP-----KETA 483
           ||   | |||   |:|||   |  ||     ||||  |  | |  |: | |       |
Db    1718 TTTPSPPPTTMTTPSPTTTPSPPTTTMTTLPPTTTSSPLTTTPLPPSITPPTFSPFSTTT 1777

Qy     484 PTTP-------------------------------------------KKLTP-------- 492
           ||||                                           : |
Db    1778 PTTPCVPLCNWTGWLDSGKPNFHKPGGDTELIGDVCGPGWAANISCRATMYPDVPIGQLG 1837

Qy     493 ----------------------------------------TTPEKLAPTTPEKPAPTT 510
                                                   | |  :  || | | | |
Db    1838 QTVVCDVSVGLICKNEDQKPGGVIPMAFCLNYEINVQCCECVTQPTTMTTTTTENPTPPT 1897

Qy     511 PEELAPTTPEEPTPT-TPEEPAPTTPKAAAPNTPKEPAPTTPKEP-----APTTPKEPAP 564
             :  ||   |||| |  :   |||          | ||  : |       ||   | |
Db    1898 TTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTP 1957
```

GenCore version 4.5
Copyright (c) 1993 - 2000 Compugen Ltd.

OM protein - protein search, using sw model

Run on: April 26, 2002, 16:32:21 ; Search time 62.75 Seconds
(without alignments)
742.061 Million cell updates/sec

Title: AA3
Perfect score: 6814
Sequence: 1 MAWKTLPIYLLLLLSVFVIQ.....ARAITTRSGQTLSKVWYNCP 1270

Scoring table: BLOSUM62
Gapop 10.0 , Gapext 0.5

Searched: 100059 seqs, 36664827 residues

Total number of hits satisfying chosen parameters: 100059

Minimum DB seq length: 0
Maximum DB seq length: 2000000000

Post-processing: Minimum Match 0%
Maximum Match 100%
Listing first 45 summaries

Database : SwissProt_39:*

Pred. No. is the number of results predicted by chance to have a score greater than or equal to the score of the result being printed, and is derived by analysis of the total score distribution.

SUMMARIES

Result No.	Score	% Query Match	Length	DB	ID	Description
1	1171	17.2	5179	1	MUC2_HUMAN	Q02817 homo sapien
2	950	13.9	1664	1	SLP1_CLOTM	Q06852 clostridium
3	798.5	11.7	1367	1	AMYH_YEAST	P08640 saccharomyc
4	651	9.6	1255	1	MUC1_HUMAN	P15941 h mucin 1 p
5	556.5	8.2	875	1	FP1_MYTED	Q25460 mytilus edu
6	555.5	8.2	2700	1	ZAN_HUMAN	Q9y493 homo sapien
7	551	8.1	620	1	EXTN_TOBAC	P13983 nicotiana t
8	533	7.8	1087	1	NFH_MOUSE	P19246 mus musculu
9	530.5	7.8	1162	1	TCNA_TRYCR	P23253 trypanosoma
10	530	7.8	865	1	CPN_DROME	Q02910 drosophila
11	518.5	7.6	872	1	FP1_MYTCO	Q25434 mytilus cor
12	509.5	7.5	662	1	MUC1_XENLA	Q05049 xenopus lae
13	503.5	7.4	1970	1	RPB1_HUMAN	P24928 homo sapien
14	498.5	7.3	1970	1	RPB1_MOUSE	P08775 mus musculu
15	497	7.3	831	1	NFH_RAT	P16884 rattus norv
16	493.5	7.2	467	1	RPB1_CRIGR	P11414 cricetulus
17	488.5	7.2	826	1	SSP2_PLAYO	Q01443 plasmodium
18	488	7.2	1020	1	NFH_HUMAN	P12036 homo sapien
19	475.5	7.0	267	1	EXTN_MAIZE	P14918 zea mays (m
20	472	6.9	2142	1	BAT2_HUMAN	P48634 homo sapien
21	471.5	6.9	5376	1	ZAN_MOUSE	O88799 mus musculu
22	468.5	6.9	634	1	HWP1_CANAL	P46593 candida alb
23	467	6.9	817	1	VRP1_YEAST	P37370 saccharomyc
24	455	6.7	1161	1	YJ9P_YEAST	P47179 saccharomyc
25	454	6.7	797	1	VGLX_HSVEB	P28968 equine herp
26	448.5	6.6	670	1	VG50_HSVI1	Q00130 ictalurid h
27	442.5	6.5	751	1	FP1_MYTGA	Q27409 mytilus gal
28	439.5	6.4	1083	1	T2D3_HUMAN	O00268 homo sapien
29	432.5	6.3	439	1	XP2_XENLA	P17437 xenopus lae
30	432	6.3	3164	1	TEGU_HSV11	P10220 herpes simp
31	430.5	6.3	3421	1	TEGU_HSVEB	P28955 equine herp
32	426	6.3	2715	1	TRX2_HUMAN	Q9umn6 homo sapien
33	424.5	6.2	1125	1	MAP4_MOUSE	P27546 mus musculu
34	424	6.2	307	1	SGS3_DROME	P02840 drosophila
35	424	6.2	1251	1	YQU3_CAEEL	Q09550 caenorhabdi
36	422.5	6.2	1794	1	YAV1_SCHPO	Q10172 schizosacch
37	422	6.2	2774	1	MAPA_RAT	P34926 rattus norv
38	421	6.2	1229	1	N121_HUMAN	Q9y2n3 homo sapien
39	420	6.2	1411	1	TCOF_HUMAN	Q13428 homo sapien
40	419.5	6.2	2476	1	ZAN_PIG	Q28983 sus scrofa
41	419	6.1	1185	1	DRPL_HUMAN	P54259 homo sapien
42	411.5	6.0	907	1	VGP3_EBV	P03200 epstein-bar
43	410	6.0	1183	1	DRPL_RAT	P54258 rattus norv
44	410	6.0	2517	1	NCR2_HUMAN	Q9y618 h nuclear r
45	407.5	6.0	3256	1	KI67_HUMAN	P46013 homo sapien

ALIGNMENTS

```
RESULT   1
MUC2_HUMAN
ID   MUC2_HUMAN     STANDARD;      PRT;  5179 AA.
AC   Q02817; Q14878;
DT   01-JUN-1994 (Rel. 29, Created)
DT   01-NOV-1997 (Rel. 35, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   MUCIN 2 PRECURSOR (INTESTINAL MUCIN 2).
GN   MUC2 OR SMUC.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Intestine;
RX   MEDLINE=94132002; PubMed=8300571;
RA   Gum J.R. Jr., Hicks J.W., Toribara N.W., Siddiki B., Kim Y.S.;
RT   "Molecular cloning of human intestinal mucin (MUC2) cDNA.
RT   Identification of the amino terminus and overall sequence similarity
RT   to prepro-von Willebrand factor.";
RL   J. Biol. Chem. 269:2440-2446(1994).
RN   [2]
RP   SEQUENCE OF 626-1895 AND 4196-5179 FROM N.A.
RC   TISSUE=Colon;
RX   MEDLINE=93016075; PubMed=1400449;
RA   Gum J.R. Jr., Hicks J.W., Toribara N.W., Rothe E.-M., Lagace R.E.,
RA   Kim Y.S.;
RT   "The human MUC2 intestinal mucin has cysteine-rich subdomains located
RT   both upstream and downstream of its central repetitive region.";
RL   J. Biol. Chem. 267:21375-21383(1992).
RN   [3]
RP   SEQUENCE OF 1343-1895 AND 4176-4195 FROM N.A.
RX   MEDLINE=91358717; PubMed=1885763;
RA   Toribara N.W., Gum J.R. Jr., Culhane P.J., Lagace R.E., Hicks J.W.,
RA   Petersen G.M., Kim Y.S.;
RT   "MUC-2 human small intestinal mucin gene structure. Repeated arrays
RT   and polymorphism.";
RL   J. Clin. Invest. 88:1005-1013(1991).
CC   -!- FUNCTION: COATS THE EPITHELIA OF THE INTESTINES, AIRWAYS, AND
CC       OTHER MUCUS MEMBRANE-CONTAINING ORGANS. THOUGHT TO PROVIDE A
CC       PROTECTIVE, LUBRICATING BARRIER AGAINST PARTICLES AND INFECTIOUS
CC       AGENTS AT MUCOSAL SURFACES.
CC   -!- SUBUNIT: MULTIMERIC.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- TISSUE SPECIFICITY: COLON, SMALL INTESTINE, COLONIC TUMORS,
CC       BRONCHUS, CERVIX AND GALL BLADDER.
CC   -!- PTM: ALL CYSTEINE RESIDUES ARE INVOLVED IN INTRACHAIN OR
CC       INTERCHAIN DISULFIDE BONDS (BY SIMILARITY).
CC   -!- POLYMORPHISM: THE NUMBER OF REPEATS IS HIGHLY POLYMORPHIC AND
CC       VARIES AMONG DIFFERENT ALLELES.
CC   -!- SIMILARITY: THE N-TERMINAL DOMAIN SHOWS SOME SIMILARITY TO THAT
CC       OF SILKWORM HEMOCYTIN.
CC   -!- SIMILARITY: CONTAINS 2 VWFC DOMAINS.
CC   -!- SIMILARITY: CONTAINS 1 C-TERMINAL CYSTINE KNOT-LIKE DOMAIN (CTCK).
CC   ---------------------------------------------------------------------
```

```
Qy     800 APKMTKETATTTEKTTESKITATTTQVTSTTTQDTTP 836
           |           |    |     ||  | |  : |
Db     704 PPVKPPPVQVPPTPTYSPPIKPPPVQVPPTPTTPSPP 740
```

Search completed: April 26, 2002, 16:31:52
Job time: 707 sec